Spring 2021

Structural Analysis of Protein-Peptide Interactions

Melody Gao
Western Washington University

Structural Analysis of Protein-Peptide Interactions

Melody Gao
Honors Program, Western Washington University
Honors Senior Capstone

Part 1: Structural Characterization and Computational Analysis of PDZ domains in *Monosiga brevicollis* (Adapted from Gao et al. 2020. Protein Sci.)

Abstract

Identification of the molecular networks that facilitated the evolution of multicellular animals from their unicellular ancestors is a fundamental problem in evolutionary cellular biology. Choanoflagellates are recognized as the closest extant non-metazoan ancestors to animals. These unicellular eukaryotes can adopt a multicellular-like "rosette" state. Therefore, they are compelling models for the study of early multicellularity. Comparative studies revealed that a number of putative human orthologs are present in choanoflagellate genomes, suggesting that a subset of these genes were necessary for the emergence of multicellularity. However, previous work is largely based on sequence alignments alone, which does not confirm structural nor functional similarity. Here, we focus on the PDZ domain, a peptide-binding domain which plays critical roles in myriad cellular signaling networks and which underwent a gene family expansion in metazoan lineages. Using a customized sequence similarity search algorithm, we identified 178 PDZ domains in the *Monosiga brevicollis* proteome. This includes 11 previously unidentified sequences, which we analyzed using Rosetta and homology modeling. To assess conservation of protein structure, we solved high resolution crystal structures of representative *M. brevicollis* PDZ domains that are homologous to human GIPC and SHANK1 PDZ domains. To assess functional conservation, we calculated binding affinities for mbGIPC, mbSHANK and mbSNX27 PDZ domains from *M. brevicollis*. We find that peptide selectivity is generally conserved between these two disparate organisms. Overall, our results provide novel insight into signaling pathways in a choanoflagellate model of primitive multicellularity.

Introduction

The events in molecular evolution that led to the origination of multicellular eukaryotes are preserved in the genomes of choanoflagellates, and they are recognized as the closest living relatives to the common ancestor of metazoans and unicellular eukaryotes.[1,2] Of particular interest in evolutionary cellular biology is the ability of choanoflagellates to adopt a primitive multicellular state, known as a rosette.[3–5] Comparative studies have revealed that several human gene families have clear orthologs in the choanoflagellate clade, and many of these orthologous genes are differentially expressed during development of the rosette.[6] Furthermore, over 350 gene families previously thought to be present only in animal lineages exist in choanoflagellate genomes.[6] Therefore, there is interest in understanding the molecular underpinnings of signaling pathway proteins in choanoflagellates, in order to gain insight into how multicellularity emerged. Investigators have identified a multitude of conserved protein structural domains and architectures in the choanoflagellate proteome that are essential to metazoan intracellular signaling systems and development. Some examples include the Notch receptor, kinases (e.g., Src family kinases, CamKII, etc.), ubiquitin ligases (e.g., Cbl), and PDZ domains, amongst others.[2,7–11] Notably, although functionally present, mechanisms of regulation can vary dramatically; while the phosphorylation-dependent regulation of Cbl is conserved in metazoans and choanoflagellates, the allosteric regulation of the SH3-SH2-Kinase module of Src family kinases is distinct.[9,12,13]

Of the shared gene families in metazoans and choanoflagellates, the PDZ domain is particularly interesting for a number of reasons. PDZ domains play key functional roles in neuronal signaling, and the intercellular

attachments that are formed during rosette development are reminiscent of neuronal synapses.[4,14–16] In addition, PDZ domains are overrepresented in the genome of choanoflagellate *Monosiga brevicollis* relative to the unicellular eukaryote *Saccharomyces pombe*.[2] Finally, PDZ domains are known to have proliferated in the metazoan lineage.[16–19] Collectively, these data suggest that PDZ domains played an important role in the evolution of multicellularity and that further characterization of PDZ domains in choanoflagellates may yield insights into molecular mechanisms that facilitated primitive multicellular development. PDZ domains were named after the first PDZ domain-containing proteins that were identified (PSD-95, Dlg1, and ZO-1).[20–24] These initially discovered PDZ domains all contain a "GLGF" amino acid sequence. This shared sequence, referred to as the GLGF-loop, or the carboxylate-binding loop, comprises a key component of the canonical PDZ domain structure. The well-conserved PDZ domain structure dictates its function in mediating protein-protein interactions. Specifically, the amide nitrogen atoms in the backbone of the GLGF-loop directly interact with the carboxylate atoms of the extreme C-terminus of a protein ligand.[25,26] PDZ domains are approximately 80-100 residues in length, and comparative analysis of hundreds of PDZ structures in the Protein Data Bank reveals a conserved structural fold, consisting of a core antiparallel b-sheet and 1-2 a-helices (**Figure 1**).[26–32]

PDZ domains are scaffolding domains that bind target proteins. In some instances, this facilitates localization of target proteins within close proximity of auxiliary enzymatic domains on the same polypeptide. In other instances, PDZ domain scaffolding activity functions to mediate protein trafficking, and impacts cellular signaling pathways. These PDZ domain interactions can be modulated by other protein-protein interaction domains on the same polypeptide, or *in trans*, by other proteins in larger macromolecular complexes.[26,33,34] An example of the scaffolding function of PDZ domains is the postsynaptic density of neurons, where multiple receptor signaling networks are brought into close physical proximity due to a number of PDZ domain-mediated interactions.[35] As mentioned previously, an expansion of the number of PDZ domain-containing genes coincided with the emergence of animal multicellularity.[17] This suggests that PDZ domains played a critical role in the evolution of multicellular animals. The human proteome contains 272 PDZ domains in a variety of protein architectures, but all PDZ domains share the same basic biochemical function of scaffolding protein-protein interactions.

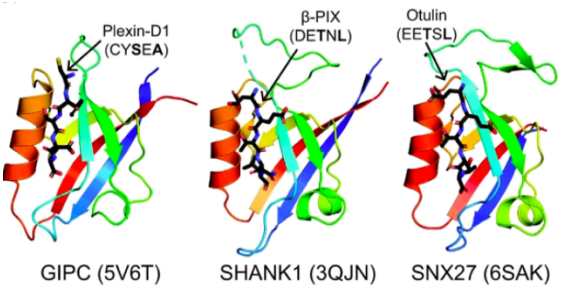

Figure 1. Conserved fold of PDZ domain structures. The human PDZ homologues of the M. brevicollis PDZ domains studied in this paper are shown in cartoon representation, colored by conserved secondary structure elements, as labeled. Bound peptides are in black stick and labeled.

Considering the importance of PDZ domains in a number of cellular processes, significant effort has been invested in characterizing their peptide-binding selectivity. These domains bind to short sequences in target proteins, often interacting with only 6 amino acid residues. In fact, the motifs of classically determined PDZ binding classes are dependent on only two residues, the extreme C-terminal residue, termed P^0 and two residues adjacent, or P^{-2}.[26] For example, Class I PDZ domains recognize the motif X-S/T-X- ϕ at the C-terminus of target proteins (where X=any amino acid and ϕ=any hydrophobic amino acid).[26] Work in the last 10+ years using high throughput techniques, e.g., phage display, peptide array, or the *hold-up assay*, has shifted this classical view of PDZ domain binding to appreciate the importance of binding interactions at non-motif residues in the peptide-binding cleft.[36–38] In addition, a number of elegant studies using directed evolution, or other protein engineering techniques, have successfully identified structural elements that determine PDZ selectivity -- often through only a small number of amino acid substitutions or post-translational modifications.[39–43]

The elucidation of PDZ binding selectivity has enabled investigators to trace the evolution of PDZ specificity throughout the tree of life, including in bacteria, yeast, and plants.[18,42,44] However, what remains to be

determined is whether or not the selectivity determinants in PDZ domains related by evolution are also conserved, despite different signaling pathways, e.g., in uni- versus multicellular organisms, or those with and without a nervous system. Previous work looking at the evolution of PDZ domains found that six amino acid positions determine lineage relationships amongst 40,000 PDZ domains in 40 proteomes and that four of these positions are in direct contact with non-motif peptide residues (P^{-1} and P^{-3}).[16] This result suggests that homologous proteins will share conserved residues in the peptide-binding cleft, including those amino acids that directly interact with residues beyond the P^0 and P^{-2} motif positions.

In order to investigate these questions on a molecular level, we crystallized and solved four total structures of two PDZ domains from the choanoflagellate, *Monosiga brevicollis*, including homologues of PDZ domains from the human proteins GIPC1 and SHANK1 (**Figure 1**). We also investigated the binding affinities of a homologue of human SNX27 (**Figure 1**). These proteins are important in postsynaptic signaling and well-conserved in *M. brevicollis*, despite over 200 million years of evolution between the last common ancestor of humans and choanoflagellates -- and the emergence of neurons.[35,45–47] Structural and binding affinity analyses confirm that the residues in the peptide-binding clefts are generally conserved in these proteins. Previous studies investigated the molecular basis of evolution, expansion, and *rewiring* in PDZ domain networks; however, here we find that for closely related PDZ domains, selectivity determinants for all residues in the binding cleft are generally conserved in evolution, despite a lack of conservation in shared target proteins.[48,49]

Results

Structural and biochemical characterization of mbGIPC PDZ

To determine if residues that directly interact with the ligand are conserved, including all of those within the peptide-binding cleft of Class I PDZ domains, we set out to characterize a number of PDZ domains from *Monosiga brevicollis* with clear homology to human PDZ proteins. We first chose to investigate the homolog of the human GAIP interacting protein, C terminus, or GIPC.[50] GIPC was first identified as an interactor of the G-alpha interacting protein (GAIP), but was quickly shown to also interact directly with G-protein coupled receptors, as well as dopamine and NMDA receptors in excitatory synapses of the central nervous system.[47,50–52] Thus, GIPC is important for both G-protein coupled receptor (GPCR) and neuronal signaling in human cells, and additional studies have shown that it broadly regulates vesicular trafficking of many transmembrane receptors via interactions with myosin VI.[30]

The presence of a GIPC homolog in *M. brevicollis* is consistent with the identification of adhesion GPCRs in choanoflagellates.[53,54] Overall, full length GIPC proteins from human and *M. brevicollis* (UniProt ID: A9VCZ3_MONBE, termed mbGIPC) share 56% sequence identity over 79% of the protein. The human GIPC protein is a Class I PDZ binding domain, as defined above. Recognition of the P^{-2} Ser/Thr residue is facilitated by hydrogen bond formation with a conserved histidine in the first position of the conserved αB helix, termed αB-1.[26] The human and choanoflagellate GIPC PDZ domains are 58% identical over 88% of the PDZ sequence, as defined by UniProt boundaries, including shared carboxylate-binding loop sequences of ALGL and conservation of the Class I-defining histidine in the αB-1 position (**Figure S1A**).

We expressed and purified mbGIPC PDZ using previously described methods, and as described in more detail in the Materials and Methods.[38,55] Briefly, we used recombinant expression in *Escherichia coli* cells, followed by affinity and size exclusion chromatography to produce purified mbGIPC PDZ protein. With protein in hand, we crystallized and solved the structure of mbGIPC PDZ to a high resolution of 1.2 Å, as described in the Materials and Methods and Supplementary Information. Overall, this structure is consistent with the conserved PDZ fold, characterized by the central five-stranded antiparallel β-sheet (βA-E) (**Figure 1**). As mentioned above, while many PDZ domains contain two α-helices (αA-B), it appears that aA is slightly strained and therefore not

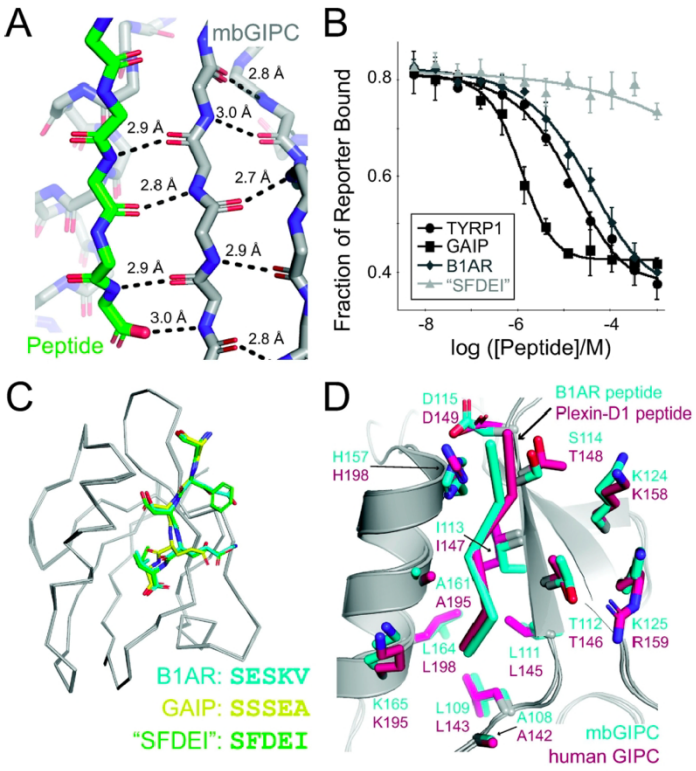

Figure 2. The crystal structure of the mbGIPC PDZ domain. (A) The interaction of mbGIPC (gray) with the C-terminal tail of a molecule related by symmetry (green), backbone atoms shown in stick, reveals a canonical PDZ-peptide interaction where the peptide forms an additional strand of an antiparallel b-sheet. Distances are labeled. (B) Average fluorescence polarization displacement isotherms are shown for mbGIPCtrunc PDZ. Titration curves correspond to the following peptides: GAIP (circles), B1AR (squares), TYRP1 (diamonds), and a decameric peptide matching the C-terminal residues of the construct, ending in "SFDEI" (triangles). Error bars indicate the standard deviation from the mean for triplicate experiments. (C) Alignment of mbGIPC PDZ domains with three separate C-terminal tail sequences (gray ribbon, RMSD = 0.21 Å for ~350 main chain atoms), with tail sequences as sticks and colored as labeled. (D) The conservation between mbGIPC (gray cartoon, with cyan side chain residues as sticks; peptide is in cyan ribbon) and human GIPC (PDB ID: 5V6B, with Plexin-D1 C-terminal peptide from 5V6T (hot pink ribbon); gray cartoon with hot pink side chain residues as sticks). Residues in the peptide-binding cleft are labeled. All stick representation is colored by heteroatom (O=red, N=blue).

fully formed in the mbGIPC structure, a characteristic that is also true of the human GIPC PDZ domain (PDB IDs: 5V6B and 5V6T).[30] The peptide ligand forms an additional strand of the central β-sheet (**Figure 2A**). Data collection and refinement statistics are in **Table S1A**.

Although we had added a peptide matching the GAIP sequence, a high affinity human GIPC PDZ target, during crystallization, we were surprised to see that our crystal structure lacked the bound peptide. Instead mbGIPC was interacting with the C-terminal tail of a molecule related by symmetry (**Figure S1B**). This is a common mode of co-crystallization for PDZ domains and ligands, e.g., in the NHERF1 PDZ bound to the cystic fibrosis transmembrane conductance regulator (CFTR) structure, as well as others.[29,56] The NHERF1 PDZ1-CFTR example is distinct from our structure, however, in that the C-terminus of mbGIPC is not a Class I PDZ-satisfying motif (sequence: KSFDEI). In our structure, which we will refer to as $mbGIPC_{SFDEI}$, we see that the P^0 Ile is accommodated by a hydrophobic pocket, as expected in Class I PDZ interactions. However, the conserved αB-1 H157 residue is forming hydrogen bonds with the Asp in the P^{-2} position (distance: 2.6 Å), as well as the Ser in the P^{-4} position (2.8 Å) (**Figure S1C**).

In order to determine if this interaction is a crystal artifact, we created a truncated mutant, $mbGIPC_{trunc}$, lacking the final 7 residues of our original construct (or K181Δ), and calculated binding affinities for human GIPC targets using fluorescence polarization. We first measured the binding affinity of $mbGIPC_{trunc}$ for a decameric fluorescent reporter peptide matching the sequence of GAIP (*F**-QGPSQSSSEA, where *F** = FITC or fluorescein isothiocyanate), calculating a $K_D = 0.29 \pm 0.02$ μM in a quadruplicate experiment (**Figure S1D**). Next, we determined the affinities of a number of human GIPC PDZ targets using competition experiments, including decameric peptides of the C-termini of GAIP (QGPSQSSSEA), tyrosinase-related protein 1 (TYRP1, sequence: KLQNPNQSVV), and the β-1 adrenergic receptor (B1AR, sequence: RPGFASESKV) (**Table 1, Figure 2B**). Experimental protocols were based on previously described methods and are described in more detail in the Materials and Methods.[38,55,57,58]

The binding affinities of $mbGIPC_{trunc}$ PDZ for human GIPC PDZ targets suggest a large degree of conservation in selectivity determinants. Specifically, the affinity of mbGIPC PDZ for GAIP is 0.23 μM, despite a BLASTP search revealing no obvious GAIP homolog in *M. brevicollis*.[59] This result also suggests a minimal

effect of the fluorescein moiety in binding. The binding affinities for TYRP1 and B1AR are ~10 and 20x worse, respectively (**Figure 2B**). These values are still relatively high to average affinity as compared to typical PDZ domain interactions, which can range from the nanomolar to hundreds of micromolar range, but are centered around 1-30 μM (11, 48, 50). Neither TYRP1 nor B1AR have clear homologues in *M. brevicollis*, according to BLASTP.[59] Notably, a competition experiment with a decameric peptide matching the C-terminal sequence of our original construct ("SFDEI" sequence: VEPKKSFDEI) revealed little to no binding, defined here as a K_i > 1000 μM (**Table 1, Figure 2B**). Thus, we concluded that the binding interaction in our original structure was a crystal artifact. In order to investigate the stereochemistry of a peptide binding interaction with mbGIPC that is not an artifact of crystallization, we mutated the final 5 residues of our original construct to those matching B1AR ($mbGIPC_{B1AR}$; C-terminal sequence: SESKV), GAIP ($mbGIPC_{GAIP}$; SSSEA), and TYRP1 ($mbGIPC_{TYRP1}$; NQSVV) (**Table S1A**). Previous work from ourselves and others suggests that the P^{-5} position is an important selectivity determinant in some PDZ domains.[37,38] However, we chose to keep this residue a lysine in our new constructs, due to crystal lattice contacts made by the lysine side chain, suggesting it may be important for crystallization (**Figure S1E**).

Table 1. Binding affinities of $mbGIPC_{trunc}$ PDZ domain.

		K_i (μM)
	Sequence	mbGIPC
GAIP	QGPSQSSSEA	0.23 ± 0.04
TYRP1	KLQNPNQSVV	3.3 ± 1.0
B1AR	RPGFASESKV	8.5 ± 4.6
mbGIPC C-term	VEPKKSFDEI	>1000

All 3 complexes successfully crystallized in the same space group as $mbGIPC_{SFDEI}$ and we determined crystal structures of $mbGIPC_{B1AR}$ and $mbGIPC_{GAIP}$. The overall conformations of these structures to each other, as well as to the $mbGIPC_{SFDEI}$ structure, were very similar, with pairwise structural alignment RMSD values ≤ 0.21 Å for ~350 main chain atoms (**Figure 2C**). We were unable to fully refine the $mbGIPC_{TYRP1}$ structure despite a successful molecular replacement solution, due to anisotropic data and relatively low resolution, compared to the others. Partial refinement (R_{work}/R_{free} = 24.3/28.9) shows clear peptide-specific density, confirming that this sequence interacts with mbGIPC in a manner that is consistent with PDZ domain peptide binding (**Figure S1F**). However, our structural analyses of the mbGIPC and human GIPC PDZ domains will be limited to the $mbGIPC_{B1AR}$ and $mbGIPC_{GAIP}$ structures.

Our mbGIPC structures share high structural similarity with the human GIPC PDZ domain. Structural alignment of main chain atoms between the $mbGIPC_{B1AR}$ and hGIPC PDZ domain (PDB ID: 5V6T) is 0.607 Å over 299 main chain atoms. This human GIPC PDZ structure was crystallized with the intracellular region of Plexin-D1 (C-terminal sequence: CYSEA) and the structures confirm that the peptide binding clefts of mbGIPC and human GIPC are very well conserved, with only two conservative substitutions (using human GIPC numbering): T148S and R159K (**Figure 2D**). We were unable to purify soluble human GIPC PDZ in our lab, despite testing multiple constructs (including using a SUMO-tag), but our data strongly suggests that the binding affinities would be similar between these domains.

Structure Characterization of mbSHANK1 PDZ

We previously compared binding affinities for another *M. brevicollis* PDZ domain, that of mbSHANK1 (UniProt ID: A9V7E4_MONBE), a protein that is homologous to human SHANK1 (**Figure S2A**).[55] In this work, we also created a homology model of mbSHANK1 PDZ using SwissModel and predicted stereochemical differences in the peptide binding pockets between these two proteins, specifically in those residues that interact with the P^{-3} position.[55] Here, we expand that investigation by presenting the crystal structure of mbSHANK1 PDZ (**Figure 3A, Table S1B**).

The protein mbSHANK1 PDZ was expressed and purified as previously described.[55] Crystallization of this protein in complex with a fluoresceinated peptide matching the C-terminus of GIRK3 (*F**-GIRK3, sequence: *F**-LPPPESESKV) is described in the Materials and Methods and Supporting Information. We collected data to a high resolution of ~2.2 Å, however phasing by molecular replacement and structure refinement proved challenging. We employed an iterative Rosetta modeling approach coupled with Phenix in order to determine a molecular replacement solution with high confidence, as described in detail in the Supplementary Information.[60–63] Our refinement difficulties were due to a large degree of anisotropy in the diffraction data. Specifically, the high-resolution limit along the a* and b* directions (2.2 Å) was substantially higher than that along the c* direction (3.4 Å). We were ultimately able to refine this model by truncating and scaling the reflections file appropriately, using the UCLA-DOE Diffraction Anisotropy Server.[64] Crystallization attempts with a fluorescent β-PIX peptide (sequence: *F**-NDPAWDETNL) were unsuccessful, despite binding mbSHANK1 PDZ with much higher affinity (K_D = 7.3 μM for mbSHANK1 PDZ, and 5.1 μM for SHANK1 PDZ), as previously reported (**Table 2**).[55] For comparison, previous attempts to get K_D values of mbSHANK1 and human SHANK1 PDZ domains with the *F**-GIRK3 peptide were incomplete, with estimates of affinities >1000 μM for each (data not shown). We were also unable to grow crystals of mbSHANK1 in the apo form or following incubation with non-fluorescent versions of either the β-PIX or GIRK3 peptides.

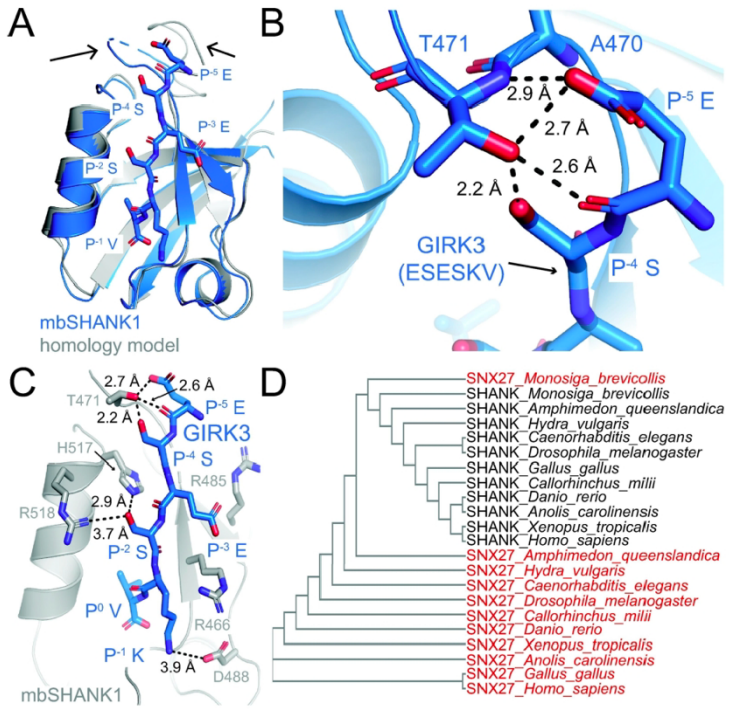

Figure 3. The crystal structure of the mbSHANK1 PDZ domain. (A) The mbSHANK1structure (blue cartoon, peptide in stick representation) is similar to a previously reported homology model (gray cartoon), RMSD = 0.668 Å over 276 main chain atoms.55 The black arrows highlight the differences in the flexible βB-βC loop. All sticks are colored by heteroatom (O=red, N=blue) and the peptide positions are labeled. (B-C) The interactions of the mbSHANK1 PDZ domain (gray cartoon, with side chains as sticks) with the F*-GIRK3 peptide (blue sticks, interchangeably referred to as "GIRK3" peptide, since the fluorescein moiety is unresolved in the crystal structure) is characteristic of PDZ-peptide interactions. Measurements between interacting residues in the peptide-binding cleft are labeled. The GIRK3 peptide is labeled, and the sequence included in the figures is in (B). There is an additional peptide residue resolved in the crystal structure, the P-6 Pro, but it does not make interactions with mbSHANK1 PDZ. (D) Phylogenetic tree showing the relationship of a number of SHANK and SNX27 PDZ domain sequences from 11 organisms. SNX27 sequences are colored red. The mbSHANK1 and mbSNX27 sequences sit at the branch point of the other SHANK and SNX27 sequences.

The crystal structure of mbSHANK1 bound to *F**- GIRK3 is structurally very similar to our previously determined homology model.[55] The overall RMSD of these two structures is 0.668 Å over 276 main chain atoms, with the largest discrepancy occurring in the flexible βB-βC loop (**Figure 3A**). In our structure, we see non-covalent interactions between T471 and the side chains of the P^{-4} Ser and P^{-5} Glu residues, as well as the P^{-5} Glu carbonyl, which may have helped to stabilize the βB-βC loop for crystallization and may explain why this complex crystallized despite a relatively low binding affinity, although six residues of the loop are disordered in our structure (**Figures 3B-C**). In addition, we see electrostatic interactions between D488 and the P^{-1} Lys, as well as H517 and R518 with the P^{-2} Ser (**Figure 3C**).

In our previous work and based on our mbSHANK1 homology model, we hypothesized that the modest increase in affinity for β-PIX by mbSHANK1 PDZ (K_i = 13 μM versus 20 μM for human SHANK1 PDZ) was

Table 2. Binding affinities of SHANK1, SNX27, mbSHANK1 and mbSNX27 PDZ domains.
SHANK1 and mbSHANK1 PDZ domain measurements are previously published, with the exception of the GIRK3 peptide.[55]

		K_i (μM)			IC_{50} (μM)
	Sequence	SHANK1	mbSHANK1	SNX27	mbSNX27
BPIX	NDPAWDETNL	20 ± 4.0	13 ± 4.7	0.26 ± 0.2	6.1 ± 4.0
GIRK3	LPPPESESKV	1070 ± 380	960 ± 160	1.7 ± 0.6	3.5 ± 0.2
mGluR1	RYKQSSSTL	90 ± 11	1020 ± 260	180 ± 60	>1000
A9UP44	EDTNQSESRL	34 ± 10	44 ± 23	0.99 ± 0.7	2.9 ± 2.7
A9UXE1	ANPIQDETAL	30 ± 5.0	72 ± 14	0.32 ± 0.2	5.9 ± 6.4
A9V7Z4	GTSLEDETAL	9.8 ± 3.1	39 ± 19	0.19 ± 0.1	1.5 ± 0.9

due to an additional arginine residue that was located near the P^{-3} position, and, we figured, positioned to interact directly with the P^{-3} Glu.[55] Interestingly, our experimental structure reveals that neither of the arginine residues in the vicinity are interacting with the P^{-3} Glu of GIRK3. However, we do see that the sequence and length of the βB-βC loop, which directly interacts with the peptide P^{-4} Ser and P^{-5} Glu residues in our structure, varies quite dramatically: the residues of the 11-residue loop for mbSHANK1 PDZ are not conserved at all with those of the 18-residue loop of human SHANK1 PDZ. It is unclear how these loops may differentially interact with the P^{-4} Asp and P^{-5} Trp of β-PIX, but otherwise, the crystal structure confirms that the peptide-binding clefts are generally conserved.[55]

When we ran our initial BLASTP search for SHANK1 PDZ homologues in *M. brevicollis*, the top two sequence hits were relatively close in sequence identity: A9V7E4_MONBE, with 34% sequence identity over 93 residues, as well as A9URU5_MONBE, with 36% sequence identity over 89 residues (domain boundaries for human SHANK1 PDZ (residues 663-757) as defined by UniProt). Sequence alignments using the full-length A9V7E4_MONBE protein and the human proteome confirmed its homology to the SHANK protein family, specifically due to the additional presence of ankyrin repeat domains, as well as SH3 and SAM domains.[45,65] Sequence alignments using the full-length A9URU5_MONBE sequence and human proteome suggested that it is a homologue of sorting nexin-27 (SNX27), with 25% sequence identity over 96% of the protein. Therefore, we will refer to A9URU5_MONBE as mbSNX27.

We were interested in the relationship between the PDZ domain sequences in these four proteins due to the similar sequence similarities between mbSHANK1 and mbSNX27 PDZ domains and human SHANK1 PDZ. Therefore, we conducted a phylogenetic tree analysis of 10 PDZ sequences for SHANK1 or SNX27 homologues in a variety of organisms, as well as the PDZ domain sequences of mbSHANK1 and mbSNX27 (**Figure 3D**). Because we see that mbSHANK1 and mbSNX27 sit at the branch point between the SNX27 and SHANK1 sequences, we expressed and purified SNX27 and mbSNX27 PDZ domains, as described in the Materials and Methods, and compared binding affinities for all four domains using fluorescence polarization to 6 decameric peptides matching the C-termini of: β-PIX, GIRK3, and mGluR1, as well as A9UP44_MONBE, A9UXE1_MONBE, and A9V724_MONBE, which were previously identified as potential *M. brevicollis* targets of mbSHANK1 (**Figures S2B-D, Table 2**).[55]

Our results reveal that overall, peptides which bind human SNX27 PDZ with relatively high affinity also bind mbSNX27 PDZ very strongly (**Table 2**). As previously reported, this is also true with human SHANK1 and mbSHANK1 PDZ domains (**Table 2**).[55] However, in all cases, the exact order of highest to lowest affinity peptides is distinct, perhaps reflective of single substitutions in the peptide binding cleft. We described the differences for SHANK1 and mbSHANK1 PDZ domains above and previously.[55] A homology model of mbSNX27 PDZ, using SNX27 PDZ as a template (PDB ID: 6SAK) contains the following substitutions at residues that may interact with the peptide (numbering based on SNX27): R58K, V61T, A83H, and R122I (**Figure S2E**). In addition, while the mGluR1 peptide binds the human PDZ domains with moderate affinity, it shows no measurable affinity for either of the *M. brevicollis* PDZ domains. Taken together, the resulting binding affinities are consistent with our central hypothesis that the target selectivity of PDZ lineages were set early in

evolution, even in proteins that appear to be closely related to each other, e.g., SHANK1 and SNX27 PDZ domains, based on overall sequence identity.

Concluding Remarks

Structural comparison of the peptide-binding clefts and peptide interactions of homologous domains from organisms related by hundreds of millions of years of evolution has the potential to provide insight into signaling networks in those species. Here, we chose to use structural biology and biochemistry to investigate two PDZ domain-containing proteins that are important in human neuronal signaling in a species of choanoflagellates, our closest non-metazoan ancestors. Many of the human targets of SHANK1 and GIPC1 are either not conserved in choanoflagellates or do not contain PDZ binding sequences.[55] However, we find that the peptide binding-cleft residues and binding affinities for human and/or choanoflagellate peptides are generally conserved in these related domains. Specifically, we see strong binding affinity correlations in SHANK1 and mbSHANK1 versus SNX27 and mbSNX27 PDZ domains, despite binding cleft substitutions in both cases.

Our structures of two unique *M. brevicollis* PDZ domains provide the first structural determination of choanoflagellate PDZ domains to our knowledge. Furthermore, our comparisons with known human PDZ domain structures, as well as homology and Rosetta modeling confirm that because the PDZ domain fold is so well conserved, it is possible to get an initial idea of a PDZ domain structure without experimental structure determination. We hypothesize that these types of analyses can be applied to PDZ domains from multiple organisms related by evolution.

Protein-protein interactions that involve PDZ domains act as critical nodes for signaling and trafficking pathways in a cell. It is clear that this is true in differentiated cells, such as those in complex multicellular organisms, as well as in single-celled organisms. Deciphering the PDZ-mediated interactions in choanoflagellates may elucidate important characteristics of the selectivity determinants and the evolution of this important peptide-binding domain. Furthermore, there are a number of proteins and protein architectures that contain PDZ domains in choanoflagellates that are not conserved in humans. Future work could investigate how these proteins, for example A9VDV9 mentioned above, act in signaling pathways in *M. brevicollis* and how this provides insight into the transition from uni- to multicellular life on Earth. Taken together, we suggest that investigating the structure-function relationship for individual domains in both uni- and multicellular organisms is an important component in building a holistic understanding of the signaling networks of an organism and in understanding the origin of multicellularity.

Materials and Methods

Protein Expression and Purification. Expression and purification of all human and *M. brevicollis* PDZ domains followed a similar protocol as previously reported for mbSHANK1 PDZ.[55] Histagged versions of the PDZ domains were inserted into the pET28a+ vector (GenScript) and expressed in *Escherichia coli* BL21 (DE3) cells. Cells were lysed using sonication and immobilized metal-affinity chromatography (5 mL HisTrap (GE Healthcare)) was used to purify proteins from the clarified supernatant. The wash buffer used was: 25 mM imidazole pH 8.5, 25 mM Tris pH 8.5, 25 mM NaCl, 10% (*v/v*) glycerol, and 0.25 mM TCEP, and elution buffer was: 400 mM imidazole pH 8.5, 25 mM Tris pH 8.5, 50 mM NaCl, 10% (*v/v*) glycerol, and 0.5 mM TCEP. With the exception of human SHANK1 and SNX27 and mbSNX27 PDZ domains, the protein was then dialyzed in dialysis buffer (same as gel filtration buffer described below), and incubated with PreScission protease to cleave off the His-tag. The cleaved protein was then purified using a second nickel column with the wash and elution buffers described above. All proteins were further purified on a Superdex S75 column, using gel filtration buffer [25 mM Tris pH 8.5, 125 mM NaCl, 10% (*w/v*) glycerol, 0.5 mM TCEP]. Proteins were concentrated using

Amicon centrifugal concentrators (3 MWCO). Concentrated proteins used in fluorescence polarization assays were flash frozen in liquid nitrogen for storage at -80°C. Proteins used for crystallization were stored at 4°C.

Proteins were quantitated with the A_{280} and the experimental extinction coefficient values of: 1490 $cm^{-1}*M^{-1}$ for all mbGIPC PDZ domains (including mbGIPC$_{SFDEI}$, mbGIPC$_{trunc}$, mbGIPC$_{B1AR}$, mbGIPC$_{GAIP}$, and mbGIPC$_{TYRP1}$), 8480 $cm^{-1}*M^{-1}$ for SHANK1 PDZ, 11000 $cm^{-1}*M^{-1}$ for mbSHANK1, 9970 $cm^{-1}*M^{-1}$ for mbSNX27 PDZ and 2980 $cm^{-1}*M^{-1}$ for SNX27 PDZ.

Crystallization, Data Collection, and Structure Determination. Prior to crystallization, all PDZ domains were dialyzed into a crystallization buffer [25 mM NaCl, 10 mM Hepes pH 7.4] for 2-4 hours. The protein concentrations used for crystallization were as follows: mbSHANK1 (6 mg/mL), mbGIPC$_{SFDEI}$ (23.4 mg/mL), mbGIPC$_{B1AR}$ (17 mg/mL), mbGIPC$_{GAIP}$ (22.8 mg/mL), and mbGIPC$_{TYRP1}$ (27.3 mg/mL). Peptide was added at a final concentration of 1 mM and incubated with protein for 1-hour prior to crystallization for *F**-GIRK3:mbSHANK1 PDZ. All initial crystallization conditions were identified using the PEG/Ion screen (Hampton Research). The crystallization conditions of crystals used for data collection were: mbGIPC$_{SFDEI}$ [100 mM ammonium tartrate dibasic pH 7.0, 12% (*w/v*) PEG 3350], mbGIPC$_{B1AR}$ [200 mM sodium malonate pH 7.0, 20% (*w/v*) PEG 3350], mbGIPC$_{GAIP}$ [4% (*v/v*) Tacsimate pH 4.0, 12% (*w/v*) PEG 3350], mbGIPC$_{TYRP1}$ [100 mM DL-Malic acid pH 7.0, 12% (*w/v*) PEG 3350], and mbSHANK1 [250 mM NaCl, 100 mM Bis-Tris pH 5.5, 32% (*w/v*) PEG 3350].

For data collection, crystals were transferred into cryoprotectant buffer. For mbSHANK1, this was well solution plus 20% (*w/v*) glycerol. For other proteins, 15% (*w/v*) glycerol was added directly to the respective PEG/Ion screen solution. The crystals were flash-cooled by plunging into liquid nitrogen. Data was collected at the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory (LBNL) on beamline 5.0.1, at l=0.977410 Å over 360°, with Df=0.25° frames and an exposure time of 0.5 s per frame. Data were processed using the XDS package (**Table S1**).[83–85] Molecular replacement was performed using Phenix with the following search models: mbGIPC (PDB ID: 5V6B, human GIPC) and mbSHANK1 (*de novo* structural model using the Robetta server and Rosetta optimization as described in the Supplemental Information).[60,86,87] Refinement was performed using Phenix, manual refinement was done using Coot, and model geometry was assessed using Molprobity and the PDB validation server.[60,86,89–93] All crystal data and refinement statistics are in **Table S1**. Additional details regarding the structure determination of mbSHANK1 is in the Supplemental Information. PDB accession codes for the structures presented here are: 6X1X (mbGIPC$_{SFDEI}$), 6X20 (mbGIPC$_{B1AR}$), 6X22 (mbGIPC$_{GAIP}$) and 6X23 (mbSHANK).

Binding assays by fluorescence polarization. Fluorescence polarization assays were performed as previously described.[25,38,55,58] Replicate experiments were performed to determine the K_D values of mbGIPC$_{trunc}$ PDZ (N=4) for the fluorescence peptide, *F**-GAIP (*FITC*-QGPSQSSSEA), and SNX27 PDZ (N=3) for the fluorescent peptides, *F**-β-PIX (*FITC*-NDPAWDETNL) and *F**-GIRK3 (*FITC*-LPPPESESKV) (**Figures S1D, S2B**). For mbGIPC$_{trunc}$ PDZ we determined a K_D value of 0.29 ± 0.02 μM for *F**-GAIP. For SNX27 we determined a K_D value of 0.022 ± 0.007 μM for *F**-β-PIX and 0.327 ± 0.135 μM for *F**-GIRK3. Limited yield of purified mbSNX27 resulted in the inability to calculate K_D values for that protein, thus, we reported IC_{50} values for our unlabeled peptides (**Figure S2D, Table 2**).

Competition experiments: The final protein concentrations for K_i experiments were equal to: 0.6 μM for mbGIPC PDZ, 0.05 μM for SNX27 PDZ, and 5 μM for mbSNX27 PDZ. For SHANK1 and mbSHANK1 PDZ K_i experiments with the GIRK3 peptide, we used 10 μM protein (based on previously-determined K_D values of

7.3 μM for mbSHANK1 PDZ, and 5.1 μM for SHANK1 PDZ).[55] Competition experiments were performed in triplicate, using the following reporter peptides at 30 nM final concentration: mbGIPC (*F**-GAIP), SNX27 (*F**-β-PIX) and mbSNX27 (*F**-β-PIX). Binding affinities for K_i experiments were determined using SOLVER and IC_{50} values using Kaleidagraph, as previously described (**Figures 2B, S2C-D)**.[38,57,58]

Literature Cited

1. Brunet T, King N (2017) The origin of animal multicellularity and cell differentiation. Dev. Cell 43:124–140.
2. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451:783–788.
3. Levin TC, Greaney AJ, Wetzel L, King N (2014) The Rosetteless gene controls development in the choanoflagellate S. rosetta. Elife 3.
4. Laundon D, Larson BT, McDonald K, King N, Burkhardt P (2019) The architecture of cell differentiation in choanoflagellates and sponge choanocytes. PLoS Biol. 17:e3000226.
5. Dayel MJ, Alegado RA, Fairclough SR, Levin TC, Nichols SA, McDonald K, King N (2011) Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. Dev. Biol. 357:73–82.
6. Richter DJ, Fozouni P, Eisen MB, King N (2018) Gene family innovation, conservation and loss on the animal stem lineage. Elife 7.
7. Shah NH, Amacher JF, Nocka LM, Kuriyan J (2018) The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases. Crit Rev Biochem Mol Biol 53:535–563.
8. del Sol A, Carbonell P (2007) The modular organization of domain structures: insights into protein-protein binding. PLoS Comput. Biol. 3:e239.
9. Amacher JF, Hobbs HT, Cantor AC, Shah L, Rivero M-J, Mulchand SA, Kuriyan J (2018) Phosphorylation control of the ubiquitin ligase Cbl is conserved in choanoflagellates. Protein Sci. 27:923–932.
10. Bhattacharyya M, Stratton MM, Going CC, McSpadden ED, Huang Y, Susa AC, Elleman A, Cao YM, Pappireddi N, Burkhardt P, et al. (2016) Molecular mechanism of activation-triggered subunit exchange in Ca(2+)/calmodulin-dependent protein kinase II. Elife 5.
11. Booth DS, King N (2020) Genome editing enables reverse genetics of multicellular development in the choanoflagellate Salpingoeca rosetta. Elife 9.
12. Li W, Young SL, King N, Miller WT (2008) Signaling properties of a non-metazoan Src kinase and the evolutionary history of Src negative regulation. J. Biol. Chem. 283:15491–15501.
13. Schultheiss KP, Suga H, Ruiz-Trillo I, Miller WT (2012) Lack of Csk-mediated negative regulation in a unicellular SRC kinase. Biochemistry 51:8267–8277.
14. Burkhardt P, Grønborg M, McDonald K, Sulur T, Wang Q, King N (2014) Evolutionary insights into premetazoan functions of the neuronal protein homer. Mol. Biol. Evol. 31:2342–2355. 15. Burkhardt P, Stegmann CM, Cooper B, Kloepper TH, Imig C, Varoqueaux F, Wahl MC, Fasshauer D (2011) Primordial neurosecretory apparatus identified in the choanoflagellate Monosiga brevicollis. Proc. Natl. Acad. Sci. USA 108:15264–15269.
16. Sakarya O, Conaco C, Egecioglu O, Solla SA, Oakley TH, Kosik KS (2010) Evolutionary expansion and specialization of the PDZ domains. Mol. Biol. Evol. 27:1058–1069.
17. Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. J. Cell Sci. 114:3219–3231.
18. Ponting CP (1997) Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci. 6:464–468.

19. Muley VY, Akhter Y, Galande S (2019) PDZ domains across the microbial world: molecular link to the proteases, stress response, and protein synthesis. Genome Biol. Evol. 11:644–659.
20. Bryant PJ, Watson KL, Justice RW, Woods DF (1993) Tumor suppressor genes encoding proteins required for cell interactions and signal transduction in Drosophila. Dev. Suppl.:239–249. 21. Woods DF, Bryant PJ (1991) The discs-large tumor suppressor gene of Drosophila encodes a guanylate kinase homolog localized at septate junctions. Cell 66:451–464.
22. Cho KO, Hunt CA, Kennedy MB (1992) The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein. Neuron 9:929–942.
23. Woods DF, Bryant PJ (1989) Molecular cloning of the lethal(1)discs large-1 oncogene of Drosophila. Dev. Biol. 134:222–235.
24. Kennedy MB (1995) Origin of PDZ (DHR, GLGF) domains. Trends Biochem. Sci. 20:350.
25. Amacher JF, Cushing PR, Bahl CD, Beck T, Madden DR (2013) Stereochemical determinants of C-terminal specificity in PDZ peptide-binding domains: a novel contribution of the carboxylatebinding loop. J. Biol. Chem. 288:5114–5126.
26. Amacher JF, Brooks L, Hampton TH, Madden DR (2020) Specificity in PDZ-peptide interaction networks: Computational analysis and review. Journal of Structural Biology: X 4:100022.
27. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85:1067–1076.
28. Slep KC (2012) Structure of the human discs large 1 PDZ2- adenomatous polyposis coli cytoskeletal polarity complex: insight into peptide engagement and PDZ clustering. PLoS One 7:e50097.
29. Elkins JM, Papagrigoriou E, Berridge G, Yang X, Phillips C, Gileadi C, Savitsky P, Doyle DA (2007) Structure of PICK1 and other PDZ domains obtained with the help of self-binding Cterminal extensions. Protein Sci. 16:683–694.
30. Shang G, Brautigam CA, Chen R, Lu D, Torres-Vázquez J, Zhang X (2017) Structure analyses reveal a regulated oligomerization mechanism of the PlexinD1/GIPC/myosin VI complex. Elife 6. 31. Stangl A, Elliott PR, Pinto-Fernandez A, Bonham S, Harrison L, Schaub A, Kutzner K, Keusekotten K, Pfluger PT, El Oualid F, et al. (2019) Regulation of the endosomal SNX27retromer by OTULIN. Nat. Commun. 10:4320.
32. Lee JH, Park H, Park SJ, Kim HJ, Eom SH (2011) The structural flexibility of the shank1 PDZ domain is important for its binding to different ligands. Biochem. Biophys. Res. Commun. 407:207–212.
33. Nourry C, Grant SGN, Borg J-P (2003) PDZ domain proteins: plug and play! Sci STKE 2003:RE7.
34. Lee H-J, Zheng JJ (2010) PDZ domains and their binding partners: structure, specificity, and modification. Cell Commun. Signal. 8:8.
35. Kim E, Sheng M (2004) PDZ domain proteins of synapses. Nat. Rev. Neurosci. 5:771–781.
36. Duhoo Y, Girault V, Turchetto J, Ramond L, Durbesson F, Fourquet P, Nominé Y, Cardoso V, Sequeira AF, Brás JLA, et al. (2019) High-Throughput Production of a New Library of Human Single and Tandem PDZ Domains Allows Quantitative PDZ-Peptide Interaction Screening Through High-Throughput Holdup Assay. Methods Mol. Biol. 2025:439–476.
37. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh J-H, Reva B, Held HA, Appleton BA, Evangelista M, Wu Y, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol. 6:e239.
38. Amacher JF, Cushing PR, Brooks L, Boisguerin P, Madden DR (2014) Stereochemical preferences modulate affinity and selectivity among five PDZ domains that bind CFTR: comparative structural and sequence analyses. Structure 22:82–93.

39. Gógl G, Biri-Kovács B, Durbesson F, Jane P, Nomine Y, Kostmann C, Bilics V, Simon M, Reményi A, Vincentelli R, et al. (2019) Rewiring of RSK-PDZ Interactome by Linear Motif Phosphorylation. J. Mol. Biol. 431:1234–1249.
40. Ferrer M, Maiolo J, Kratz P, Jackowski JL, Murphy DJ, Delagrave S, Inglese J (2005) Directed evolution of PDZ variants to generate high-affinity detection reagents. Protein Eng Des Sel 18:165–173.
41. Huang J, Koide A, Makabe K, Koide S (2008) Design of protein function leaps by directed domain interface evolution. Proc. Natl. Acad. Sci. USA 105:6578–6583.
42. Kim J, Kim I, Yang J-S, Shin Y-E, Hwang J, Park S, Choi YS, Kim S (2012) Rewiring of PDZ domain-ligand interaction network contributed to eukaryotic evolution. PLoS Genet. 8:e1002510. 43. Skelton NJ, Koehler MFT, Zobel K, Wong WL, Yeh S, Pisabarro MT, Yin JP, Lasky LA, Sidhu SS (2003) Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. J. Biol. Chem. 278:7645–7654.
44. Teyra J, Ernst A, Singer A, Sicheri F, Sidhu SS (2020) Comprehensive analysis of all evolutionary paths between two divergent PDZ domain specificities. Protein Sci. 29:433–442.
45. Sheng M, Kim E (2000) The Shank family of scaffold proteins. J. Cell Sci. 113 (Pt 11):1851–1856.
46. Sheng M, Kim E (2011) The postsynaptic organization of synapses. Cold Spring Harb. Perspect. Biol. 3.
47. Yi Z, Petralia RS, Fu Z, Swanwick CC, Wang Y-X, Prybylowski K, Sans N, Vicini S, Wenthold RJ (2007) The role of the PDZ protein GIPC in regulating NMDA receptor trafficking. J. Neurosci. 27:11663–11675.
48. Ernst A, Gfeller D, Kan Z, Seshagiri S, Kim PM, Bader GD, Sidhu SS (2010) Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. Mol. Biosyst. 6:1782–1790.
49. te Velthuis AJW, Sakalis PA, Fowler DA, Bagowski CP (2011) Genome-wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network. PLoS One 6:e16047.
50. De Vries L, Lou X, Zhao G, Zheng B, Farquhar MG (1998) GIPC, a PDZ domain containing protein, interacts specifically with the C terminus of RGS-GAIP. Proc. Natl. Acad. Sci. USA 95:12340–12345.
51. Jeanneteau F, Guillin O, Diaz J, Griffon N, Sokoloff P (2004) GIPC recruits GAIP (RGS19) to attenuate dopamine D2 receptor signaling. Mol. Biol. Cell 15:4926–4937.
52. Jeanneteau F, Diaz J, Sokoloff P, Griffon N (2004) Interactions of GIPC with dopamine D2, D3 but not D4 receptors define a novel mode of regulation of G protein-coupled receptors. Mol. Biol. Cell 15:696–705.
53. Krishnan A, Schiöth HB (2015) The role of G protein-coupled receptors in the early evolution of neurotransmission and the nervous system. J. Exp. Biol. 218:562–571.
54. Krishnan A, Almén MS, Fredriksson R, Schiöth HB (2012) The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. PLoS One 7:e29817.
55. Valgardson J, Cosbey R, Houser P, Rupp M, Van Bronkhorst R, Lee M, Jagodzinski F, Amacher JF (2019) MotifAnalyzer-PDZ: A computational program to investigate the evolution of PDZ-binding target specificity. Protein Sci. 28:2127–2143.
56. Karthikeyan S, Leung T, Ladias JA (2001) Structural basis of the Na+/H+ exchanger regulatory factor PDZ1 interaction with the carboxyl-terminal region of the cystic fibrosis transmembrane conductance regulator. J. Biol. Chem. 276:19683–19686.
57. Vouilleme L, Cushing PR, Volkmer R, Madden DR, Boisguerin P (2010) Engineering peptide inhibitors to overcome PDZ binding promiscuity. Angew. Chem. Int. Ed. Engl. 49:9912–9916.
58. Cushing PR, Fellows A, Villone D, Boisguérin P, Madden DR (2008) The relative binding affinities of PDZ partners for CFTR: a biochemical basis for efficient endocytic recycling. Biochemistry 47:10084–10098.

59. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215:403–410.
60. McCoy AJ (2007) Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr. Sect. D, Biol. Crystallogr. 63:32–41.
61. DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, et al. (2011) Improved molecular replacement by density- and energyguided protein structure optimization. Nature 473:540–543.
62. DiMaio F, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D (2013) Improved lowresolution crystallographic refinement with Phenix and Rosetta. Nat. Methods 10:1102–1104. 63. DiMaio F, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits T, Cheng Y, Baker D (2015) Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with densityguided iterative local refinement. Nat. Methods 12:361–365.
64. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. USA 103:8060–8065.
65. Monteiro P, Feng G (2017) SHANK proteins: roles at the synapse and in autism spectrum disorder. Nat. Rev. Neurosci. 18:147–157.
66. Chaptal V, Kilburg A, Flot D, Wiseman B, Aghajari N, Jault J-M, Falson P (2016) Two different centered monoclinic crystals of the E. coli outer-membrane protein OmpF originate from the same building block. Biochim. Biophys. Acta 1858:326–332.
67. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40:D71-5.
68. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47:D506–D515.
69. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011:bar009.
70. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211-5.
71. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, et al. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 42:W252-8.
72. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46:W296–W303.
73. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a webbased environment for protein structure homology modelling. Bioinformatics 22:195–201.
74. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2009) Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. 4:1–13.
39
75. Ford AS, Weitzner BD, Bahl CD (2020) Integration of the Rosetta suite with the python software stack via reproducible packaging and core programming interfaces for distributed simulation. Protein Sci. 29:43–51.
76. Lau Y-TK, Baytshtok V, Howard TA, Fiala BM, Johnson JM, Carter LP, Baker D, Lima CD, Bahl CD (2018) Discovery and engineering of enhanced SUMO protease enzymes. J. Biol. Chem. 293:13224–13233.

77. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) High-resolution comparative modeling with RosettaCM. Structure 21:1735–1742.
78. Nivón LG, Moretti R, Baker D (2013) A Pareto-optimal refinement method for protein design scaffolds. PLoS One 8:e59004.
79. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. 23:47–55.
80. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al. (2017) The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 13:3031–3048.
81. Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. 179:125–142.
82. Ivarsson Y, Arnold R, McLaughlin M, Nim S, Joshi R, Ray D, Liu B, Teyra J, Pawson T, Moffat J, et al. (2014) Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. Proc. Natl. Acad. Sci. USA 111:2542–2547.
83. Kabsch W (2010) XDS. Acta Crystallogr. Sect. D, Biol. Crystallogr. 66:125–132.
84. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. Acta Crystallogr. Sect. D, Biol. Crystallogr. 66:133–144.
85. Kabsch W (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. J Appl Crystallogr 26:795–800.
86. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. Sect. D, Biol. Crystallogr. 66:213–221.
87. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32:W526-31.
88. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ, Adams PD (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr. Sect. D, Biol. Crystallogr. 64:61–69.
89. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr. Sect. D, Biol. Crystallogr. 66:486–501.
90. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D, Biol. Crystallogr. 66:12–21.
91. Laskowski RA, Moss DS, Thornton JM (1993) Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol. 231:1049–1067.
92. Vaguine AA, Richelle J, Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr. Sect. D, Biol. Crystallogr. 55:191–205.
93. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 10:980.
94. NCBI Resource Coordinators (2017) Database resources of the national center for biotechnology information. Nucleic Acids Res. 45:D12–D17.

Supplemental Tables and Figures:

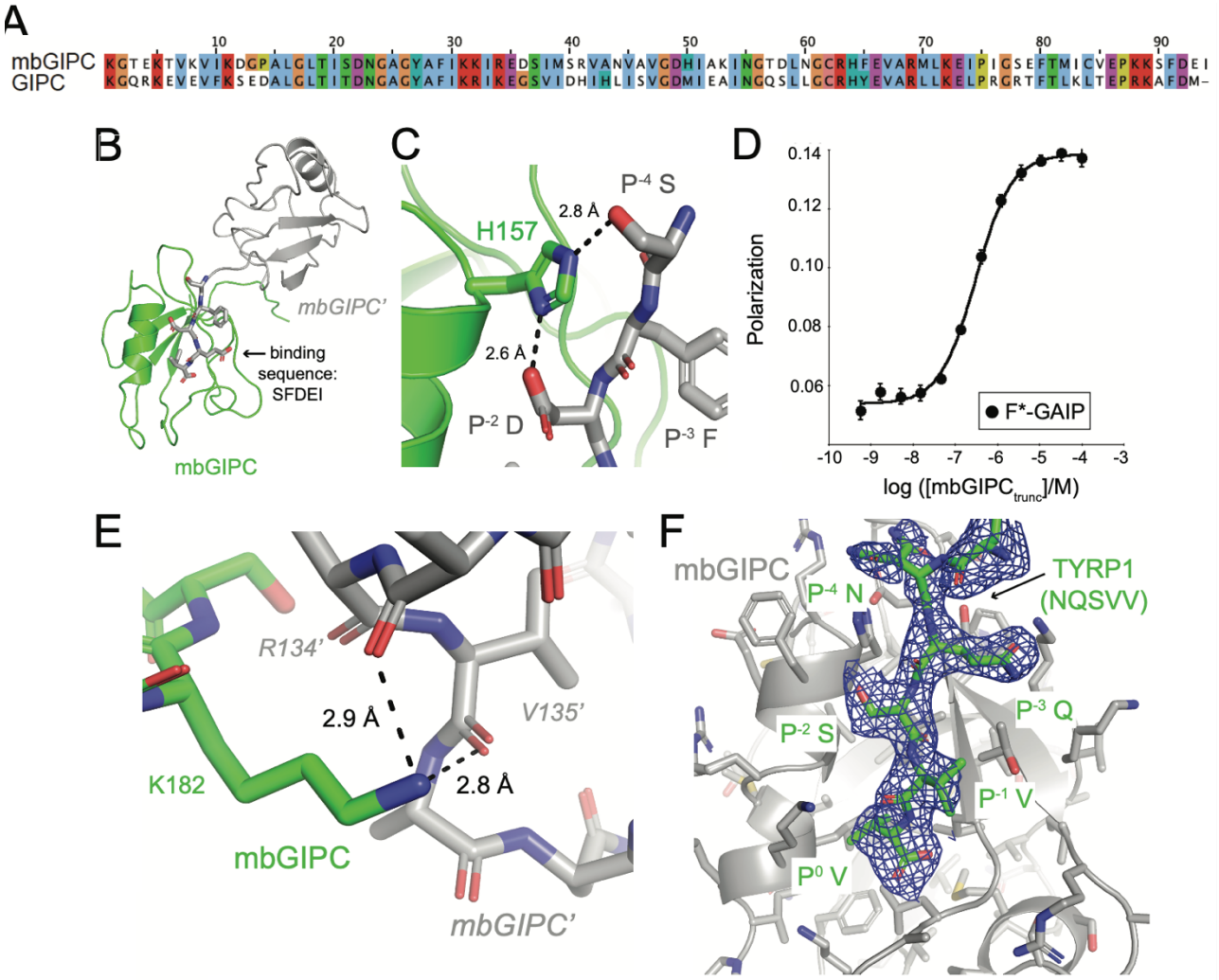

Figure S1. Crystal contacts in mbGIPCSFDEI structure and electron density for peptide in mbGIPCTYRP1. (A) Sequence alignment of human GIPC and mbGIPC PDZ domains. (B) The spatial relationship between mbGIPC (green cartoon) PDZ and a molecule related by symmetry (gray cartoon, mbGIPC') is shown. The last 5 residues of the mbGIPC' C-terminal tail (sequence SFDEI) are shown in gray stick. (C) The binding of mbGIPC (green cartoon, H157 side chain in stick representation) to the C-terminal tail of mbGIPC' (gray sticks and labeled) is unconventional and a crystal artifact. (D) The average fluorescence polarization isotherm is shown for mbGIPCtrunc PDZ and the F*-GAIP peptide (at 30 nM), including the standard deviation for each data point. This experiment was performed in quadruplicate and the calculated KD = 0.29 ± 0.02 μM. (E) The K182 residue in mbGIPCSFDEI PDZ (green sticks) makes crystal lattice contacts with the main chain carbonyl atoms of R134 and V135 in a molecule related by symmetry, mbGIPC' (gray sticks). For this reason, we chose to keep a lysine in this position in our other mbGIPC constructs (mbGIPCB1AR, mbGIPCGAIP, and mbGIPCTYRP1). (F) Final refinement of mbGIPCTYRP1 proved challenging, so we did not deposit this structure in the Protein Data Bank. However, there is clear electron density for the C-terminal sequence, NQSVV (2Fo-FC map in blue mesh and contoured at 1s), and it is consistent with mbGIPC binding to the other C-terminal sequences (GAIP and B1AR). Here, mbGIPC is in gray cartoon with side chain sticks in gray. The TYRP1 sequence is in green stick representation and labeled. For all, sticks are colored by heteroatom (O=red, N=blue).

Figure S2

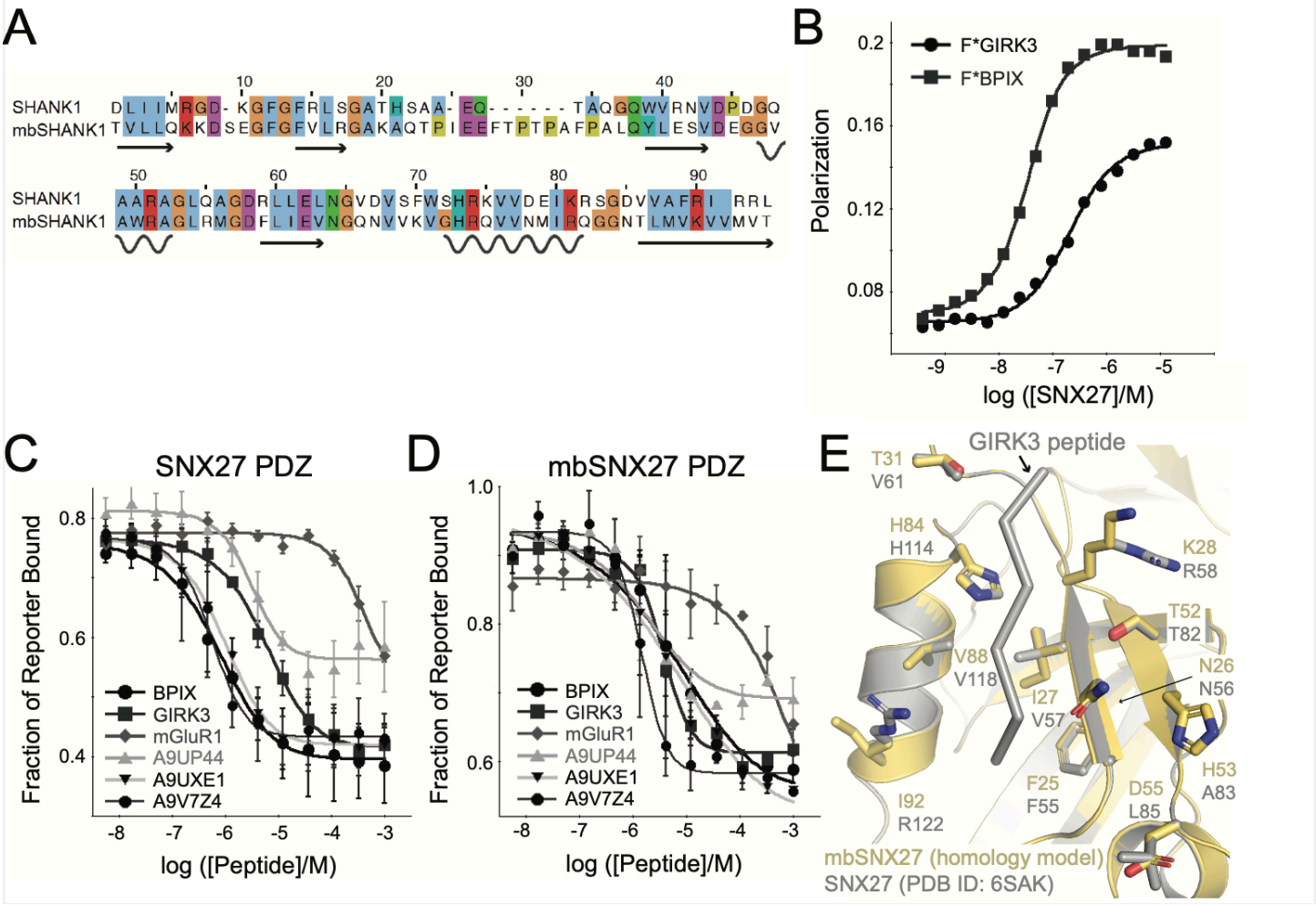

Figure S2. Structural and biochemical characterization of mbSNX27. (A) Sequence alignment of human SHANK1 and mbSHANK1 PDZ domains. Secondary structure elements are labeled by arrows (b-strands) and wavy lines (a-helices). (B) Example fluorescence polarization isotherms are shown for SNX27 PDZ and the F*-b-PIX (black squares) and F*-GIRK3 peptides (black circles, both reporter peptides at 30 nM). (C-D) Average fluorescence polarization displacement isotherms are shown for SNX27 PDZ (C) and mbSNX27 PDZ (D). Titration curves are shown for the following peptides: b-PIX (circles), GIRK3 (squares), and mGluR1 (diamonds), or choanoflagellate proteins A9UP44 (triangles), A9UXE1 (upside-down triangles), and A9V7Z4 (gray circles). Error bars indicate standard deviation from the mean for triplicate experiments. The reporter peptide used in both experiments was F*-β-PIX. (E) The conservation between mbSNX27 (gold cartoon, with side chain residues as sticks) and human SNX27 (PDB ID: 6SAK, gray cartoon with side chain residues as sticks) is shown. The peptide (GIRK3 sequence: ESESKV) is from an additional human SNX27 structure (3QE1) and is shown as gray ribbon and labeled. The RMSD value between the human SNX27 structures is 0.377 Å over 302 main chain atoms. The mbSNX27 PDZ structure was made using SwissModel with 6SAK as a template. Residues in the peptide-binding cleft are labeled. All stick representation is colored by heteroatom (O=red, N=blue).

Table S1. Data Collection and Refinement Statistics.

(A)

	mbGIPC$_{SFDEI}$	**mbGIPC$_{B1AR}$**	**mbGIPC$_{GAIP}$**	**mbGIPC$_{TYRP1}$**[g]
Data Collection				
Space Group	*P* 2$_1$ 2$_1$ 2$_1$ (19)	*P* 2$_1$ 2$_1$ 2$_1$ (19)	*P* 2$_1$ 2$_1$ 2$_1$ (19)	*P* 2$_1$ 2$_1$ 2$_1$ (19)
Unit cell dimensions				
a,b,c (Å)	29.71, 35.30, 82.77	29.79, 34.84, 81.60	29.73, 35.03, 82.55	29.34, 35.36, 80.66
α,β,γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution[a] (Å)	41.4-1.2 (1.27-1.2)	40.8-1.4 (1.49-1.4)	41.3-1.4 (1.49-1.4)	40.3-2.2 (2.34 -2.2)
R_{sym}[b] (%)	2.9 (11.5)	2.9 (128.2)	2.2 (112.2)	5.4 (79.4)
I/σ_I	31.90 (5.15)	21.58 (1.34)	22.08 (1.68)	19.97 (2.15)
Completeness (%)	91.4 (52.9)	99.9 (99.5)	96.8 (93.5)	99.9 (99.7)
Refinement				
Total # of reflections	27,011	17,306	16,970	4,553
Reflections in the test set	1,335	870	856	220
R_{work}[d]/R_{free}[e]	16.7/18.9	20.8/22.6	18.7/22.1	24.3/28.9
Number of atoms:				
Protein	765	748	723	721
Water	189	102	124	5
Ramachandran plot[f] (%)	100.0/0/0	100.0/0/0	100.0/0/0	96.9/3.1/0
B_{av} (Å^2)				
Protein	12.3	26.9	23.2	54.5
Bond length RMSD (Å)	0.005	0.005	0.006	0.010
Bond angle RMSD (°)	0.818	0.729	0.867	1.115

[a]Values in parentheses are for data in the highest-resolution shell.
[b]$R_{sym} = \Sigma_h\Sigma_i |I(h) - I_i(h)| / \Sigma_h\Sigma_i I_i(h)$, where $I_i(h)$ and $I(h)$ values are the *i*-th and mean measurements of the intensity of reflection *h*.
[c]SigAno = $|F(+) - F(-)| / \sigma$
[d]$R_{work} = \Sigma ||F_{obs}|_h - |F_{calc}||_h / \Sigma |F_{obs}|_h$, $h \in$ {working set}
[e]R_{free} is calculated as R_{work} for the reflections $h \in$ {test set}
[f]Favored/allowed/outliers
[g]Refinement not complete. Structure not deposited in Protein Data Bank.

(B)

	mbSHANK1	mbDLG-2 (1)	mbDLG-2 (2)	mbDLG-3
Data Collection				
Space Group	*P* 2_1 2_1 2_1 (19)	*I* 2 (5)	*P* 2_1 2_1 2_1 (19)	*C* 2 (5)
Unit cell dimensions				
a,b,c (Å)	33.07, 40.69, 78.86	37.15, 35.26, 63.49	35.09, 56.11, 57.92	76.38, 44.75, 28.61
α,β,γ (°)	90, 90, 90	90, 102.3, 90	90, 90, 90	90, 97.95, 90
Resolution[a] (Å)	36.2-2.15 (2.51-2.15)	34.7-1.7 (1.81-1.7)	40.3-1.3 (1.38-1.3)	38.5-1.3 (1.38-1.3)
R_{sym}[b] (%)	16 (155.0)	7.3 (54.2)	3.3 (36.0)	2.1 (13.6)
I/σ_I	12.45 (2.03)	11.39 (2.23)	24.04 (2.48)	34.04 (6.25)
Completeness (%)	99.9 (99.7)	98.4 (97.3)	99.4 (96.5)	96.5 (83.5)
Refinement				
Total # of reflections	3,941	8,833	28,680	23,070
Reflections in the test set	189	428	1,432	1,126
R_{work}[d]/R_{free}[e]	20.6/25.3	18.2/23.8	19.5/20.7	15.8/16.9
Number of atoms:				
Protein	742	710	720	816
Water	23	100	123	166
Ramachandran plot[f] (%)	96.7/2.2/1.1	100.0/0/0	100.0/0/0	100.0/0/0
B_{av} (Å^2)				
Protein	38.0	22.3	24.3	14.9
Bond length RMSD (Å)	0.010	0.006	0.005	0.005
Bond angle RMSD (°)	1.344	0.743	0.769	0.839

[a]Values in parentheses are for data in the highest-resolution shell.
[b]$R_{sym} = \Sigma_h\Sigma_i |I(h) - I_i(h)| / \Sigma_h\Sigma_i I_i(h)$, where $I_i(h)$ and $I(h)$ values are the *i*-th and mean measurements of the intensity of reflection *h*.
[c]SigAno = $|F(+) - F(-)| / \sigma$
[d]$R_{work} = \Sigma ||F_{obs}|_h - |F_{calc}||_h / \Sigma |F_{obs}|_h$, $h \in$ {working set}
[e]R_{free} is calculated as R_{work} for the reflections $h \in$ {test set}
[f]Favored/allowed/outliers

Part 2: Structural and binding studies of Class A sortase of *Streptococcus agalaticae*

Abstract (adapted from Piper et al. In review. JBC.)

Gram-positive bacteria contain sortase enzymes on their cell surfaces that catalyze transpeptidation reactions critical for proper cellular function. *In vitro*, sortases are used in sortase-mediated ligation (SML) reactions for a variety of protein engineering applications. Historically, sortase A from *Staphylococcus aureus* (saSrtA) has been the enzyme of choice for SML reactions. However, the stringent specificity of saSrtA for the sequence motif LPXTG limits its uses. Here we investigate the contribution of this β7– β8 loop, located between the catalytic cysteine and arginine residues and immediately adjacent to the target binding cleft, by designing and testing chimeric sortase enzymes. Our chimeras utilize natural sequence variation of Class A sortases from four species engineered into the SrtA sequence from *Streptococcus agalaticae* (sagSrtA). We crystallize and solve the first known full-length structure of sagSrtA and are working on solving a sagSrtA chimera and peptide-bound structures. These studies provide the foundation for a deeper understanding of sortase target selectivity and can expand the sortase toolbox for future SML applications.

Introduction (adapted from Piper et al. In review. JBC.)

Sortases are cysteine transpeptidase enzymes that gram-positive bacteria use to covalently attach proteins to their cell wall for various functions, including to assemble pili or display virulence factors.[1–3] There are 6 recognized classes of sortase enzymes (classes A-F), with *in vivo* roles ranging from general purpose or "housekeeping" functions (classes A and E), to more specific roles such as the construction of the bacterial pilus (Class C).[1,4] These enzymes recognize a cell wall sorting signal (CWSS) on the outer membrane of gram-positive bacteria.[1,5] For Class A sortases, the CWSS is the sequence LPXTG.[1,5] Using previously published numbering (L=P4, P=P3, X=P2, T=P1, and G=P1'), P4, P3 and/or P1' of this motif vary amongst different classes.[5] Following target recognition, a His-Cys-Arg catalytic triad facilitates a transpeptidation reaction whereby the CWSS is first cleaved between the P1 and P1' residues via nucleophilic attack by the catalytic Cys, resulting in a thioester linkage with the P1 position of the CWSS. Resolution of this acyl-enzyme intermediate is then achieved by nucleophilic attack by an amino group displayed on the cell wall building block lipid II, or in the case of pilus formation, displayed on a separate protein subunit.[1,3,5,6] The final result is formation of a new amide linkage, with the portion of the substrate N-terminal to the CWSS now covalently attached at its C-terminus to the amine nucleophile ligation partner.

The ability to cleave a substrate sequence and subsequently ligate a second component (for example a protein or synthetic peptide derivative) make sortases an attractive tool for protein

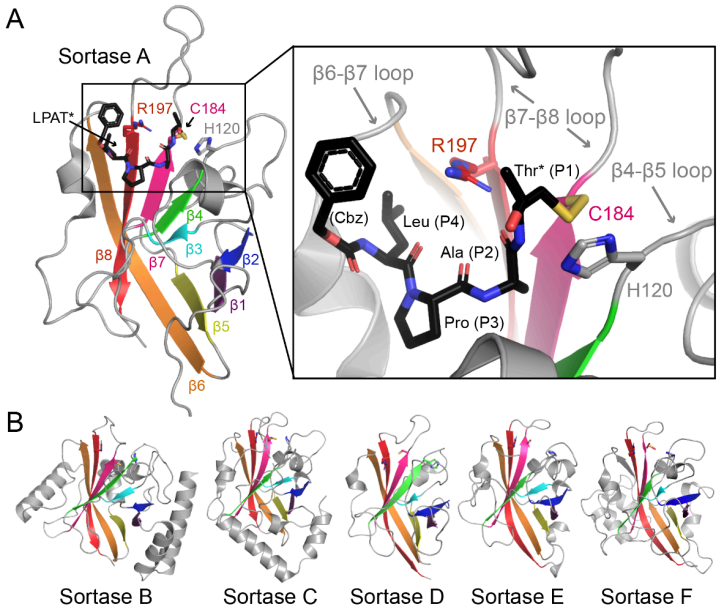

Figure 1. The sortase-fold is conserved in all classes of bacterial sortases. (A) The peptide-bound structure of *S. aureus* SrtA (saSrtA) is shown in cartoon representation, with β-strands colored and labeled (PDB ID: 2KID) [17]. The side chains of the catalytic residues (H120, C184, and R197) are shown as sticks, colored by heteroatom (O=red, N=blue, S=yellow), and labeled. The disulfide -linked peptide analog, Cbz-LPAT*, where Cbz is a carbobenzyloxy protecting group and T* is (2*R*,3*S*)-3-amino-4-mercapto-2-butanol, is shown as black sticks and colored by heteroatom [17]. A zoomed-in version of the active site is shown in the black box, with features indicated as in (A). The variable loops are labeled and indicated by gray arrows. (B) The overall sortase-fold is well conserved in proteins of different classes. Here, structures for Class B (PDB ID 1NG5), Class C (3O0P), Class D (2LN7), Class E (5CUW), and Class F (5UUS) sortases are in cartoon, with conserved β-strands colored as in (A), highlighting the 8-stranded sortase-fold. The conserved catalytic triad is shown in sticks (and colored by heteroatom) for all.

engineering efforts, commonly called sortase-mediated ligation (SML) or *sortagging*.[3] Sortase A from *Staphylococcus aureus* (saSrtA) was the first of these enzymes discovered and continues to see widespread use for *in vitro* SML experiments.[1,7] Recent years have seen notable improvements in SML technology, including strategies for limiting the reversibility of the ligation reaction, and the development of saSrtA variants with dramatically improved catalytic efficiency.[3,8,9] However, as a consequence of the narrow substrate selectivity of saSrtA[10], the majority of SML examples rely on the combination of one ligation partner displaying an LPXTG motif near its C-terminus with another possessing one or more N-terminal glycines. This restricted substrate scope can be advantageous, for example in the use of SML for labeling specific polypeptides in complex mixtures, but it also represents a limitation for certain applications (PMID: 33290621).[11,12] Highlighting this point, an increasing number of studies have demonstrated that the use of naturally occurring sortases or engineered sortases with altered substrate selectivity offer distinct advantages such as reducing the necessity for point mutations in protein semisynthesis applications[12], enabling the labeling of endogenous proteins that do not naturally contain the LPXTG motif[11,13], and allowing labeling of multiple sites within the same protein target.[11,14] Thus, the engineering and discovery of sortases with altered substrate profiles, along with a better understanding of the biochemical basis for sortase substrate selectivity, represent important areas for the continued development of SML technology.

With respect to substrate recognition, previous mutagenesis and structural studies of various sortases provide a wealth of knowledge about initial ligand recognition and subsequent cleavage (thioesterification), as well as nucleophile recognition and mechanistic details of peptide ligation (transpeptidation).[1,2,10] Specifically, the catalytic residues of all native sortases identified to date are (using saSrtA numbering unless specified otherwise): His120 (general acid/base), Cys184 (nucleophile, acyl-enzyme intermediate), and Arg197 (transition state stabilization) (**Figure 1A**).[1,10] Additionally, directed evolution studies have identified mutations (P94R/D160N/D165A/K190E/K196T) that are together able to boost the catalytic efficiency of saSrtA by 120-fold.[8] Of these 5 mutations, several are located in two of the three structurally conserved loops in Class A sortases located near the peptide-binding cleft: those between the β4, β5 strands (β4-β5 loop), the β6, β7 strands (β6-β7 loop, where D165A occurs), and the β7, β8 strands (β7-β8 loop, where K190E and K196T are located). Notably, while the increase in enzyme activity afforded by these mutations included a 3.6-fold increase in k_{cat}, the effect was dominated by a 33-fold decrease in K_M, suggesting these loop residues may be important in CWSS recognition.[8]

Additional evidence for the role of loop residues has been obtained from more targeted directed evolution and mutagenesis studies. For example, it has been demonstrated that the β6-β7 loop of saSrtA directly confers specificity at P4 of the recognition motif (LPXTG), and residues other than leucine (L) can be accommodated using sortases with mutations in the β6-β7 loop.[11,12,15] Indeed, substitution of the β6-β7 loop residues from saSrtB into the saSrtA enzyme alters substrate recognition to that of a sortase B protein (NPQTN).[16] Turning to the β7-β8 loop, the NMR structure of saSrtA covalently bound to a modified LPAT* peptide mimetic revealed a non-covalent interaction between W194 in saSrtA and the Thr residue in P1 (LPXTG).[17,18] Mutation of W194 in saSrtA decreased the reaction rate, although it was not essential to catalysis.[18] Taken together, these past studies reveal that sequence variation within sortase loops directly affects both activity and selectivity for target ligands. Furthermore, conservation of the closed eight-stranded β-barrel core in all sortase A-F structures that have been reported to date suggests that these principles may apply to non-Class A sortases as well (**Figure 1B**).[2]

The published structure of S. pneumoniae SrtA (spSrtA, PDB ID: 4O8L) is of a domain swap dimer. In its dimer form, spSrtA is inactive thus its structure does not give insight to the active conformation of Class A sortases. Looking in the Protein Database, we found Streptococcus agalaticae (sagSrtA, PDB ID: 3RCC) crystallized as a dodecamer with six monomeric units forming a ring and two rings stacked together. With the knowledge that sagSrtA has previously been crystallized, it forms the basis for our structural investigation of chimeric sagSrtA variants and co-crystallization structures of the LPATXG substrate with sagSrtA.

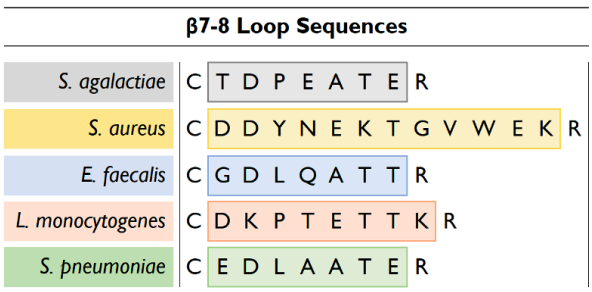

Figure 2. β7–β8 loop sequences of chimeric variants.

Results

Loop-swapped β7–β8 chimeric enzymes selectivity for P1'

Previous research in the lab has shown that the broaden substrate scope can be attributed to the role that β7–β8 loop playing a key role in the differing substrate preferences.[19] Moreover, variations in β7–β8 loop sequences can substantially impact overall enzyme activity, affording chimeric sortases that outperform their wild-type counterpart i*n vitro*.[19] Using the full-length sequence of SrtA in *Streptococcus agalaticae* as the base, four chimeric enzymes were created using previous tested β7–β8 loops to confirm trends seen before. These chimeras include β7–β8 loop residues from *Staphylococcus aureus* (sagSrtA$_{aureus}$), *Enterococcus faecalis* (sagSrtA$_{faecalis}$), *Listeria monocytogenes* (sagSrtA$_{monocytogenes}$), and *Streptococcus pneumoniae* (sagSrtA$_{pneumoniae}$) (**Figure 2)**. Enzymatic assays were preformed using LPAT**G**G, LPAT**A**G, and LPAT**S**G substrates to obtain the relative activity.

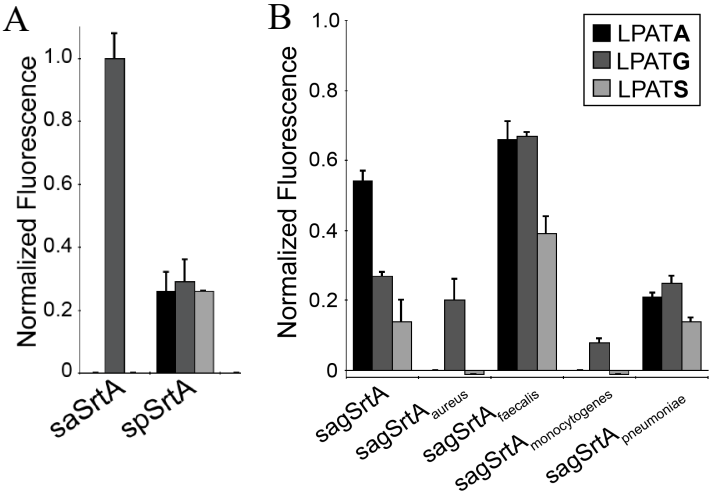

Figure 3. (A) Substrate selectivity profiles for saSrtA and spSrtA. (B) Substrate selectivity profiles for wild-type sagSrtA and chimeric sagSrtA variants. Substrate cleavage monitored via an increase in fluorescence at 420 nm from reactions of fluorophore-quencher probes with the generic structure Abz-LPATXG-K(Dnp) (LPAT**X**) in the presence of hydroxylamine. Bar graphs represent mean normalized fluorescence (± standard deviation) from at least three independent experiments.

Within the 2 hr time frame of our assay, wild-type sagSrtA is more promiscuous then saSrtA, able to react with all three substrates. This promiscuity is similar to spSrtA with a key difference in that sagSrtA cleaves LPATAG two-times more efficiently. When the *S. aureus* β7–β8 loop is swapped in, we see that sagSrtA$_{aureus}$ mimics saSrtA's selectivity for the glycine in the P1' position with an efficiency that is one-fifth of saSrtA. sagSrtA$_{faecalis}$ has an increase in efficiency with all three substrates, predominantly in the A- and G-containing peptides. The sagSrtA$_{monocytogenes}$ fundamentally mimics saSrtA being selective only for the G-containing peptide but is seen to have even less efficiency than sagSrtA$_{aureus}$. Finally, sagSrtA$_{pneumoniae}$ shows similar efficiency and promiscuity to spSrtA for all peptides except for the S-containing, being only half as efficient (**Figure 3)**. All chimeric enzymes followed promiscuity and selectivity trends seen when similar chimeric enzymes were created using the *Streptococcus pneumonia* SrtA as the base structure.[19]

Table 1. Substrate selectivity profiles for saSrtA, sagSrtA, sagSrtA$_{trunc}$ and spySrtA.

	G	S	A
saSrtA	1.00	0.00	0.00
sagSrtA	0.27	0.14	0.54
sagSrtA$_{trunc}$	0.02	-0.02	0.02
spySrtA	0.97	0.51	1.05

Full-Length Structure of Streptococcus agalaticae SrtA

The published structure of sagSrtA is of a truncated version of the enzyme, which removes the last nine residues from the C-terminus tail.[20] The effects on activity due to this truncation was investigated via fluorescence assays and revealed that sagSrtA$_{trunc}$ is inactive for all three substrates (**Table 1**). C-terminal tail is important in the enzymes structure and its ability to cleave and ligate. To investigate this difference, we expressed, purified, crystallized and solved the structure of full length sagSrtA to a high resolution of 1.4 Å. Structural alignment of sagSrtA$_{trunc}$ (PDB ID: 3RCC) to full-length sagSrtA showed significant deviations between the structures with an RMSD of 0.857 Å over 711 main chain atoms. The β6–β7 and β7–β8 loops, both important in substrate recognition

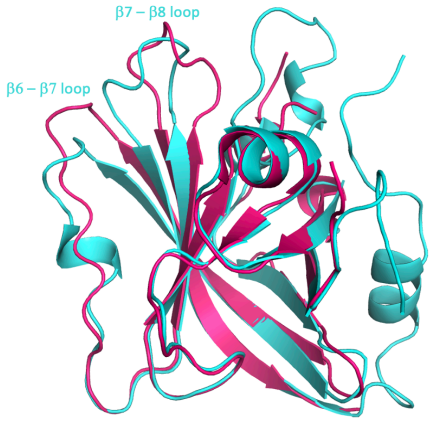

Figure 4. Structural alignment of sagSrtA$_{trunc}$ (pink) and full-length sagSrtA (aqua).

and cleavage, showed the greatest amounts of movement compared to other regions (**Figure 4**). This difference in structure we hypothesize is due to the non-endogenous dodecameric interactions that the sagSrtA$_{trunc}$ enzyme was crystallized in causing shifts in structure to pack into the oligomeric form.

Streptococcus agalaticae swap aureus SrtA Structure

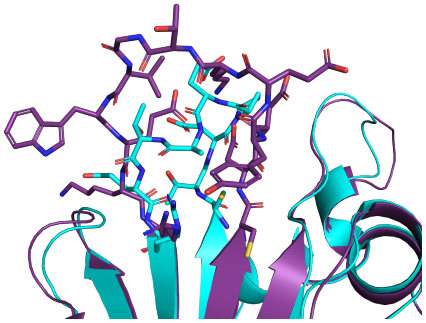

Figure 5. Structural alignment of sagSrtA (aqua) and sagSrtA$_{aureus}$ (purple). RMSD of 0.485 Å over 867 main chain atoms.

Upon obtaining the full length sagSrtA structure, we went on to crystallize the chimeric loop swaps that we had created for enzymatic assays. We were only able to crystallize the sagSrtA$_{aureus}$ even though the other loops are of more similar lengths than the S. aureus β7–β8 loop. While solving the structure we realized that the loop sequences was actually incorrect as the asparagine was omitted (**Figure 2, 5**). This enzyme was also used in the enzymatic assays and will need to be redone with the correct loop sequence. This mistake did tell us that by removing the asparagine from the loop, it does not affect the selectivity of sagSrtA. It still followed the same trend that we expected.

Co-Crystallization

With a full-length structure obtained, we went on to co-crystallize sagSrtA with our three substrates. To do so, the enzyme is first inactivated by mutating the catalytic cystine in position 206 to an alanine via site-directed mutagenesis. With the correct plasmid in hand, the inactive enzyme was expressed, purified and crystalized. G-, S- and A-containing peptides were added to a final concentration of 1 mM and incubated for an hour at room temperature prior to crystallization. The LPATGG peptide contains a fluorescein fluorophore producing yellow crystals that confirmed peptide and protein interaction. Crystals looped from these co-crystallization setups have been sent to the ALS for data collection and currently structures are being solved.

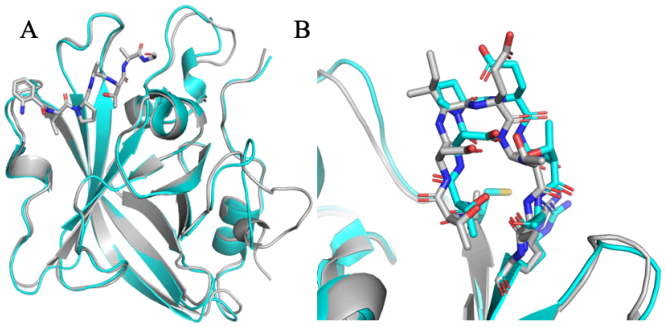

Figure 6. (A) Structural alignment of sagSrtA and spySrtA. (B) Zoomed in β7-β8 loop.

Preliminary strucutral analysis using *Streptococcus pyogenes* SrtA (spySrtA) bound to LPATAG shows a high structural alignment with an RMSD of 0.492 Å over 855 main chain atoms. Looking specifically at the loops involved in substrate recogntion and selectivity, we see minimal movement in the β6–β7 and β7–β8 loops with the greatest deviations between the structures in the β4–β5 loop. Enzymatic assays of spySrtA indicate a three-fold increase in activity for glycine and serine in the P1' position while a 2-fold increase for the alanine compared to sagSrtA (**Table 1**). Examining the β7–β8 loop sequence, there is only a one residue difference between the two. The third residue following the catalytic cystine in sagSrtA is a proline while in spySrtA is an isoleucine (**Figure 6**). The rigidity of proline effects the flexibility of the loop thus causing the decrease in activity seen in sagSrtA.

Looking more closely at LPATAG's interaction with sagSrtA, the leucine in position P4 fits nicely into a hydrophobic pocket. This pocket most likely would not be able to accommodate large hydrophobic residues due to the size of the cavity (**Figure 7A**). The P1' position of the peptide falls in a hydrophobic ridge which is also restrictive of large hydrophobic residues (**Figure 7B**). The S-containing peptide is accommodated in this ridge as the hydroxyl group is able to point out of the pocket. Finally, there are multiple hydrogen bonds made between

sagSrtA and positions P1, P2 and P3 of the peptide, specifically with the catalytic arginine and the backbone of the peptide (**Figure 7C**).

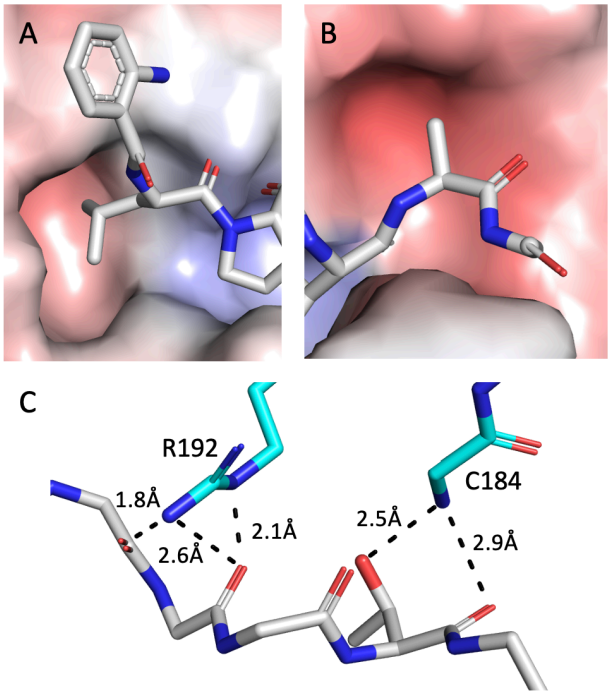

Figure 7. Electrostatic surface of sagSrtA with LPATAG. (A) Hydrophobic pocket where leucine sits. (B) Hydrophobic ridge that restricts large hydrophobic residues. (C) Hydrogen bonding of sagSrtA with peptide backbone and side chain.

Conclusion

Target sequence recognition for Class A sortases is not all rigidly selective like *S. aureus* SrtA for a P1' glycine but it seen to be able to accommodate many residues.[14,21–23] Building off previous work in the lab, we look to understand the fundamentals of sortase substate recognition through structural analysis. *S. agalaticae* SrtA as the launching off point of these studies due to a published structure. From our studies, we also solved the first full-length structure of sagSrtA and discovered that the published structure was a crystal artifact. We also solved the structure of a chimeric variant, sagSrtA$_{aureus}$, but due to a wrong loop sequence, our finding cannot be confirmed just yet but the asparagine that was omitted does not seem to be a key residue in substrate recognition. sagSrtA aligned to a high degree to spySrtA allowing for preliminary structural analysis of peptide bound structure. Similar to spySrtA, sagSrtA also exhibited: (1) a hydrophobic pocket that accommodates the P4 position, (2) multiple hydrogen bond interactions and (3) a hydrophobic ridge limiting the residues compatible in the P1' position. We also found that the proline within the β7−β8 loop sequence of sagSrtA greatly hindered its ability to efficiently cleave its substrate. This work also has implications for the further development of sortase-mediated ligation (SML) as a protein engineering tool.[3,24]

The development of these new sortase/substrate pairs has exciting consequences for SML engineering efforts: (1) it increases options for dual-labeling single proteins or multiplexed labeling of multiple proteins in the same systems[25–27], and (2) it may reduce the need to mutate naturally occurring protein sequences in order to render their termini compatible with SML. Developing a deeper understanding of how residues in these loops affect substrate selectivity in all sortase classes may enable dramatic expansion of the sortase "toolbox", potentially allowing the development of ligases that are tailored to the needs of specific protein targets while also limiting off-target effects.

Experimental Procedures (adapted from Gao et al. and Piper et al.)

Protein expression and purification. Wild-type spSrtA and saSrtA proteins were expressed and purified as previously described.[28] All other constructs, including chimeric and mutant proteins, were purchased from Genscript in the pET28a(+) vector. In general, protein expression and purification protocols were very similar to those previously described.[28] Briefly, plasmids were transformed into *Escherichia coli* BL21 (DE3) competent cells and grown in LB media, with protein induction at OD_{600} 0.6-0.8 using 0.15 M IPTG for 18-20 h at 18°C.

Following cell harvest in lysis buffer [0.05 M Tris pH 7.5, 0.15 M NaCl, 0.0005 M ethylenediaminetetraacetic acid (EDTA)], the protein was purified using a 5 mL HisTrap HP column (GE Life Sciences, now Cytiva), using wash [0.05 M Tris pH 7.5, 0.15 M NaCl, 0.02 M Imidazole pH 7.5, 0.001 M TCEP] and elution [wash buffer, with 0.3 M Imidazole pH 7.5] buffers. Size exclusion chromatography (SEC) was conducted using a HiLoad 16/600 Superdex 200 column (GE Life Sciences, now Cytiva) in SEC running buffer [0.5 M Tris pH 7.5, 0.15 M NaCl, 0.001 M TCEP]. Purified protein corresponding to the monomeric peak was concentrated using an Amicon Ultra-15 Centrifugal Filter Unit (10,000 NWML) and analyzed by SDS-PAGE. Protein not immediately used was flash frozen in SEC running buffer and stored at -80°C.

Peptide synthesis. Detailed synthetic procedures are provided in the Supplemental Data. Briefly, all peptides were synthesized via manual Fmoc solid phase peptide synthesis (SPPS). Peptides were synthesized either individually or in tandem using Fmoc Rink amide MBHA resin or Synphase lantern solid supports. All other materials, including suitably protected Fmoc amino acids, and reagents for coupling, deprotection, and resin cleavage were obtained from commercial sources and used without further purification. All peptides were purified using RP-HPLC and their identities were confirmed via ESI-MS. Prior to use in sortase-catalyzed transacylation reactions, each purified peptide was prepared as a concentrated stock solution in DMSO and/or H_2O.

Fluorescence Assay for Sortase Activity. Reactions were performed in a Costar round-bottom, black 96-well plate at a 100 μL reaction volume under the following conditions: 5 μM sortase, 50 μM peptide substrate, and 5 mM hydroxylamine nucleophile. All reactions contained 10% (*v/v*) 10x sortase reaction buffer (500 mM Tris pH 7.5, 1500 mM NaCl, and 100 mM $CaCl_2$). Reactions also contained residual DMSO from the peptide stock solutions (0.5-1.5% (*v/v*), with the exception of the Phe- and Val-containing peptides at 5%). The peptides containing phenylalanine or valine required 5% (*v/v*) DMSO for solubility under the reaction conditions. 1 mM TCEP was also included in reactions utilizing the Abz-LPATCG-K(Dnp) substrate. Reactions were initiated by the addition of the sortase enzyme, which were prepared as 10x stock solutions in 50 mM Tris pH 7.5, 150 mM NaCl, and 1 mM TCEP. Microplates were analyzed using a Biotek Synergy H1 plate reader. The fluorescence intensity of each well was measured at 2-min time intervals over a 2-hr period at room temperature (l_{ex} = 320 nm, l_{em} = 420 nm, and detector gain = 75). All reactions were performed in triplicate. For each substrate sequence, the background fluorescence of the intact peptide in the absence of enzyme was subtracted from the observed experimental data. Background-corrected fluorescence data was then normalized to the fluorescence intensity of a benchmark reaction between wild-type saSrtA and Abz-LPATGG-K(Dnp).

Crystallization, Data Collection, and Structure Determination. The protein concentrations used for crystallization was approximately 15 mg/ml. Peptide was added at a final concentration of 1 mM and incubated with protein for 1-hour prior to crystallization. All initial crystallization conditions were identified using the PEGRx screen (Hampton Research). The crystallization conditions of crystals used for data collection were: sagSrtA [100 mM MES monohydrate pH 6.0, 20% (*v/v*) 2-propanol, 20% (*w/v)* PEG monomethyl ether 2000] and sagSrtA:LPATGG [80-120 mM sodium acetate trihydrate, 26-32% (*w/v*) PEG 1500].

For data collection, all co-crystallization crystals grown from lab made well solutions were transferred into cryoprotectant buffer made of well solution plus 20% (*w/v*) glycerol. For full-length sagSrtA, 15% (*w/v*) glycerol was added directly to the respective PEGRx screen solution. The crystals were flash-cooled by plunging into liquid nitrogen. Data was collected at the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory (LBNL) on beamline 5.0.1, at l=0.977410 Å over 360°, with Df=0.25° frames and an exposure time of 0.5 s per frame. Molecular replacement was performed using Phenix with the following search model: sagSrtA (PDB ID: 3FN5, spySrtA). Refinement was performed using Phenix, manual refinement was done using Coot, and model geometry was assessed using Molprobity.

Literature Cited:

1. Spirig, T., Weiner, E. M. & Clubb, R. T. Sortase enzymes in Gram-positive bacteria. *Mol. Microbiol.* **82,** 1044–1059 (2011).
2. Jacobitz, A. W., Kattke, M. D., Wereszczynski, J. & Clubb, R. T. Sortase transpeptidases: structural biology and catalytic mechanism. *Adv. Protein Chem. Struct. Biol.* **109,** 223–264 (2017).
3. Antos, J. M., Truttmann, M. C. & Ploegh, H. L. Recent advances in sortase-catalyzed ligation methodology. *Curr. Opin. Struct. Biol.* **38,** 111–118 (2016).
4. Di Girolamo, S. *et al.* Characterization of the housekeeping sortase from the human pathogen Propionibacterium acnes: first investigation of a class F sortase. *Biochem. J.* **476,** 665–682 (2019).
5. Bradshaw, W. J. *et al.* Molecular features of the sortase enzyme family. *FEBS J.* **282,** 2097–2114 (2015).
6. Marraffini, L. A., DeDent, A. C. & Schneewind, O. Sortases and the Art of Anchoring Proteins to the Envelopes of Gram-Positive Bacteria. *Microbiol. Mol. Biol. Rev.* **70,** 192–221 (2006).

7. Mazmanian, S. K., Liu, G., Ton-That, H. & Schneewind, O. Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall. *Science* **285,** 760–763 (1999).
8. Chen, I., Dorr, B. M. & Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. USA* **108,** 11399–11404 (2011).
9. Freund, C. & Schwarzer, D. Engineered sortases in peptide and protein chemistry. *Chembiochem* (2020). doi:10.1002/cbic.202000745
10. Clancy, K. W., Melvin, J. A. & McCafferty, D. G. Sortase transpeptidases: insights into mechanism, substrate specificity, and inhibition. *Biopolymers* **94,** 385–396 (2010).
11. Bentley, M. L., Lamb, E. C. & McCafferty, D. G. Mutagenesis studies of substrate recognition and catalysis in the sortase A transpeptidase from Staphylococcus aureus. *J. Biol. Chem.* **283,** 14762–14771 (2008).
12. Schmohl, L., Bierlmeier, J., Gerth, F., Freund, C. & Schwarzer, D. Engineering sortase A by screening a second-generation library using phage display. *J Pept Sci* **23,** 631–635 (2017).
13. Kang, C.-Y. *et al.* Functional analysis of Clostridium difficile sortase B reveals key residues for catalytic activity and substrate specificity. *J. Biol. Chem.* **295,** 3734–3745 (2020).
14. Kruger, R. G. *et al.* Analysis of the substrate specificity of the Staphylococcus aureus sortase transpeptidase SrtA. *Biochemistry* **43,** 1541–1551 (2004).
15. Piotukh, K. *et al.* Directed evolution of sortase A mutants with altered substrate selectivity profiles. *J. Am. Chem. Soc.* **133,** 17536–17539 (2011).
16. Bentley, M. L., Gaweska, H., Kielec, J. M. & McCafferty, D. G. Engineering the substrate specificity of Staphylococcus aureus Sortase A. The beta6/beta7 loop from SrtB confers NPQTN recognition to SrtA. *J. Biol. Chem.* **282,** 6571–6581 (2007).
17. Suree, N. *et al.* The structure of the Staphylococcus aureus sortase-substrate complex reveals how the universally conserved LPXTG sorting signal is recognized. *J. Biol. Chem.* **284,** 24465–24477 (2009).
18. Ton-That, H., Mazmanian, S. K., Alksne, L. & Schneewind, O. Anchoring of surface proteins to the cell wall of Staphylococcus aureus. Cysteine 184 and histidine 120 of sortase form a thiolate-imidazolium ion pair for catalysis. *J. Biol. Chem.* **277,** 7447–7452 (2002).
19. Piper, I. M. *et al.* A second specificity-determining loop in Class A sortases: Biochemical characterization of natural sequence variation in chimeric SrtA enzymes. *BioRxiv* (2021). doi:10.1101/2021.03.27.437355
20. Khare, B. *et al.* Structural differences between the Streptococcus agalactiae housekeeping and pilus-specific sortases: SrtA and SrtC1. *PLoS One* **6,** e22995 (2011).
21. Antos, J. M. *et al.* Site-specific N- and C-terminal labeling of a single polypeptide using sortases of different specificity. *J. Am. Chem. Soc.* **131,** 10800–10801 (2009).
22. Race, P. R. *et al.* Crystal structure of Streptococcus pyogenes sortase A: implications for sortase mechanism. *J. Biol. Chem.* **284,** 6924–6933 (2009).
23. Raeeszadeh-Sarmazdeh, M., Parthasarathy, R. & Boder, E. T. Site-specific immobilization of protein layers on gold surfaces via orthogonal sortases. *Colloids Surf. B, Biointerfaces* **128,** 457–463 (2015).
24. Li, J. *et al.* Optimization of sortase A ligation for flexible engineering of complex protein systems. *J. Biol. Chem.* **295,** 2664–2675 (2020).
25. Dorr, B. M., Ham, H. O., An, C., Chaikof, E. L. & Liu, D. R. Reprogramming the specificity of sortase enzymes. *Proc. Natl. Acad. Sci. USA* **111,** 13343–13348 (2014).
26. Hess, G. T., Guimaraes, C. P., Spooner, E., Ploegh, H. L. & Belcher, A. M. Orthogonal labeling of M13 minor capsid proteins with DNA to self-assemble end-to-end multiphage structures. *ACS Synth. Biol.* **2,** 490–496 (2013).
27. Wuethrich, I. *et al.* Site-specific chemoenzymatic labeling of aerolysin enables the identification of new aerolysin receptors. *PLoS One* **9,** e109883 (2014).
28. Bordoli, L. *et al.* Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* **4,** 1–13 (2009).